

Making collider endorsement count

There is broad backing for a US bid to build the International Linear Collider.

A multidisciplinary panel of senior scientists last week endorsed strong US participation in the construction of the International Linear Collider (ILC), the accelerator project that particle physicists see as their top priority. In a report from the National Academy of Sciences (see page 1094), the panel called on the US government to bid to host the ILC.

The Department of Energy and the National Science Foundation, the two agencies that support high-energy physics in the United States, should take the panel's advice and develop such a bid. Unlike previous high-energy physics projects, this one will be truly global from the start. Such international projects need strong backing at the most senior levels of government if they are to progress.

The arguments for building the ILC have been well rehearsed (see *Nature* **426**, 105; 2003). The machine would mark a logical progression from the Large Hadron Collider (LHC), which is now under construction at CERN, the European particle-physics laboratory near Geneva, Switzerland. It would enable physicists to engage in the detailed study of particles, such as the Higgs boson, that they hope to find with the LHC. A US bid to host the ILC would confirm the nation's commitment to particle physics and lend valuable impetus to the collider project. Germany, Japan and CERN itself may also bid to host the project.

High-energy physics is an expensive business, and it is likely that, after the ILC, the basic approach of studying it by building ever-larger particle accelerators will need to be revisited. However, the ILC will attain clearly defined scientific goals at a level of cost — somewhere between \$5 billion and \$10 billion — that can be borne by a well-planned international effort.

The academy panel contains several eminent non-physicists, such as Harold Shapiro, an economist from Princeton University who chaired it, and biologist Harold Varmus, president of the Memorial Sloan-Kettering Cancer Center in New York and former director of the National Institutes of Health. Their endorsement of the project reflects a consensus within the wider scientific community

that particle physics, although expensive and esoteric, is of fundamental importance and worthy of support.

The effort to translate that scientific consensus into political backing for the construction project is only just beginning. So far, those involved with the ILC have done a reasonably good job of maintaining the global nature of the project and ensuring that the physics community remains united behind it. There have been some exceptions to this pattern: the insistence of CERN, for example, that decisions about the siting of the ILC be delayed until an accelerator technology it is trying to develop is ready, strikes some in the community as unnecessary and self-serving.

Although physics from the LHC should feed into the design of the ILC, an international panel chaired by Barry Barish of the California Institute of Technology selected an accelerator technology for the design two years ago, and upwards of \$100 million has already been spent on preparatory research. It is time to start talking about sites.

"The effort to translate the scientific consensus into political backing for the construction project is only just beginning."

However, such a discussion requires an expression of willingness from candidate nations to host the facility, and probably to shoulder at least half of its construction costs. Two years ago, the Office of Science at the Department of Energy identified the ILC as its main medium-term priority. But as with ITER, its top priority, there has been no sign of any willingness from the White House to allocate extra funds for the project.

In the end, political backing for the deal is likely to come from friends of Fermilab, the laboratory near Chicago where it might be built. Fermilab has firmly hitched its wagon to the ILC (see *Nature* **435**, 713; 2005) and is looking to its supporters, such as House speaker Dennis Hastert (Republican, Illinois), to get the project on to the political agenda. Armed with this latest report, that task becomes a bit easier. ■

Bioethics at the bench

Bioethicists should be close — but not too close — to the lab action.

What are the rights of parents and fetuses when genetic tests reveal a 50% likelihood of a disabling illness by the time a child reaches adulthood? Should patients be required to give active consent if their health data are to be used in the future for unforeseeable biomedical research? How should different viewpoints on religion, philosophy and the natural and social sciences be brought to bear in setting regulations?

These are the sorts of questions that bioethicists deal with. Some researchers, however, pay little heed to bioethics, or regard it as a potential impediment to their work. Concerns about the relationship between bioethics and science are as old as the field itself.

Studies into the ethical, legal and social implications (ELSI) of the human genome project, funded mainly by the National Institutes of Health (NIH), were responsible for the rapid expansion of academic bioethics in the United States. But this diverse community has attracted criticism from researchers for being too remote from real science — and from public-interest groups for being too close to it. The latter want to know how it can present a convincing critique of genomics, given its reliance on NIH institutes and investigators.

"The professional field of bioethics has a great deal to say about



many fascinating things, but people in this profession rarely say 'no,' political scientist Langdon Winner of Rensselaer Polytechnic Institute in New York state told the House science committee in April 2003. "Indeed, there is a tendency for career-conscious social scientists and humanists to become a little too cosy with researchers in science and engineering, telling them exactly what they want to hear."

After the completion of the human genome project, the ELSI programme at the NIH evolved to direct significant support to four centres of excellence. The National Human Genome Research Institute, which provides most of the funding, says it hopes that these centres will each develop a sufficiently strong identity to support the growth of the next generation of ELSI researchers.

Several of these bioethics centres of excellence say they are trying to establish closer links with more laboratory researchers (see page 1104), and that their bioethicists will make themselves more readily available to provide rapid advice to faculty members. They see this as a way of fostering more ethical thinking in the design and execution of research, long before they would otherwise think of bioethics.

Researchers have to seek ethical advice willingly and then be prepared to act on it. There is no enforcement mechanism for bioethics

or sanctions; and, unlike Institutional Review Boards, bioethics advice rarely carries any weight in law. But that should not dissuade more ethicists from getting involved with the practical applications of their expertise.

Moreover, this kind of advice may turn out to serve a useful promotional function for bioethics itself. One part of this is simply to increase awareness of ethics; it is telling that half of the enquiries during the pilot phase of the first of these projects, at Stanford University in California, simply required lab researchers to be reminded of existing ethical guidelines. But much of the advice is provided in confidence, raising issues of transparency in the event that research practices violate ethical norms.

Notwithstanding concerns about excessive cosiness with scientists, an academic bioethics community that is too aloof won't help science or ethics. As bioethicists strive to be heard beyond their own community, and for their advice to be valued, these models provide one useful way forwards. ■

"Several of these bioethics centres of excellence say they are trying to establish closer links with more laboratory researchers."

Machine readability

A publishing initiative seems ready to make text mining simpler.

For many years Tim Berners-Lee, the inventor of the World Wide Web, has dreamed of machines being able to help humans use his creation. This would enable not only sophisticated search tools to hunt for words or phrases, but also for other engines to hunt for meanings and patterns. This 'semantic web' is being pieced together gradually. The latest step forward brings users of the scientific literature closer to that dream by enhancing computer access to the full text of the scientific literature.

Many scientists are used to the idea of data mining: the ability to plunder all the available databases to search not only for relevant nuggets, but also for unexpected combinations of data that reveal — or at least hint at — relationships and mechanisms. They are not so used to the analogous function of mining texts.

But some researchers have made a start. Biologists, for example, have developed software that explores open 'text bases', especially the PubMed database. They scan many publications in order to discover relationships based on phrases or sentences that, when analysed in combination, cumulatively link one object (such as a disease) to another (such as a molecule). At the University of California, Berkeley, the BioText project is being used to explore apoptosis, for example (<http://biotext.berkeley.edu>). At the University of Illinois in Chicago, the Arrowsmith software explores the causes of disease (http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html). And at the European Bioinformatics Institute near Cambridge, UK, the EBIMed retrieval engine explores protein-protein interactions (<http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp>).

But publishers have yet to develop a standard annotation of their

content that allows computers access to the full text. Earlier this month, the Nature Publishing Group launched a preliminary proposal for such a standard. The proposal is not a commercial product but rather a potential service for the community. It is open for comment and is not intended to provide a competitive advantage to us: on the contrary, it will only succeed if adopted by other publishers.

The proposal is the Open Text Mining Interface (OTMI), which was first presented at the Life Sciences Conference and Expo in Boston earlier this month. A description and examples can be found at http://blogs.nature.com/wp/nascent/2006/04/open_text_mining_interface_1.html. The proposal would make coded text freely available to all. If all publishers were to adopt this or some similar standard, the entire literature would become accessible for mining.

How does this proposal relate to publishers' various business models? 'Author pays' publishers would be able to use this approach to machine readability and help users find their content more easily. 'Subscriber pays' publishers would follow the Nature Publishing Group in making this version of the full text freely explorable by machines but unreadable by humans. (Charging for machine access across diverse publishers' firewalls would effectively make machine text-mining impossible.) The OTMI approach to encryption is to jumble up sentences, retaining semantic relationships as far as possible.

Critics will point out that this limits the machine readability too; for example, some proximity searching becomes impossible. But the subscriber-pays model is strongly supported in the marketplace. OTMI represents a potential compromise between business needs and open access. *Nature* and its publishers welcome feedback about this initiative, which should be sent either to nature@nature.com or to the above-mentioned blog. ■

"If all publishers were to adopt this or some similar standard, the entire literature would become accessible for mining."

RESEARCH HIGHLIGHTS

COSMOLOGY

The Universe bounces back

Phys. Rev. Lett. **96**, 141301 (2006)

Will physics fit through the pinhole of the Big Bang — the first instants of space and time, when everything fitted into a volume smaller than a quark? This is the challenge for theories of quantum gravity, which pursue the elusive union of quantum mechanics and general relativity.

Abhay Ashtekar and colleagues of Pennsylvania State University argue that one of the most promising theories, known as loop quantum gravity, can preserve the predictability of physics through this singularity in space-time. They say that loop quantum gravity produces a quantum bridge between two large, classical universes: one expanding, the other contracting. Thus the Big Bang could have been a big bounce, as a previously shrinking universe reversed its contraction.

CANCER BIOLOGY

Reading the runes

Proc. Natl Acad. Sci. USA **103**, 5923–5928 (2006)

Data from thousands of patients are needed to identify the genes that influence the outcome of cancer, say researchers.

Early prediction of a cancer's potential for metastasis or relapse would improve treatment. But lists of 'predictive' genes, whose level of expression is thought to correlate with clinical outcome, have been unreliable. Lists from different studies have few genes in common.

Eytan Domany and his colleagues at the Weizmann Institute in Israel devised a mathematical method called probably approximately correct (PAC) sorting to assess the problem. They calculate that data from thousands of patients, rather than the hundreds used in studies so far, are needed to compensate statistically for the weak correlation between individual gene activity and disease outcome.

OCEANOLOGY

Ivan the terrible

Geophys. Res. Lett. **33**, L07607 (2006)

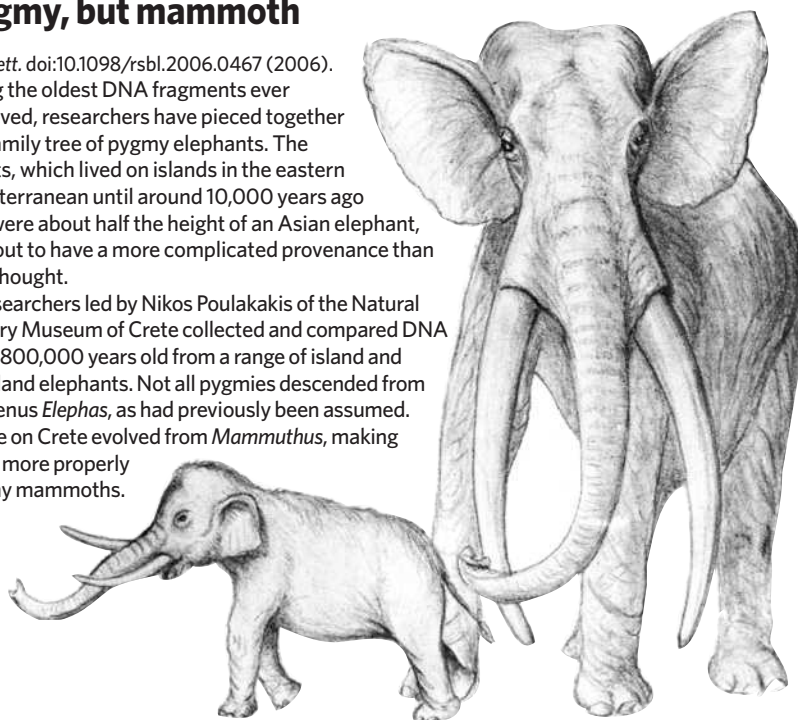
Might hurricanes trigger tsunamis? This is the question posed by William Teague and his colleagues of the US Naval Research Laboratory at the Stennis Space Center in Mississippi, on finding that Hurricane Ivan (pictured) shifted

Pygmy, but mammoth

Biol. Lett. doi:10.1098/rsbl.2006.0467 (2006).

Using the oldest DNA fragments ever retrieved, researchers have pieced together the family tree of pygmy elephants. The beasts, which lived on islands in the eastern Mediterranean until around 10,000 years ago and were about half the height of an Asian elephant, turn out to have a more complicated provenance than was thought.

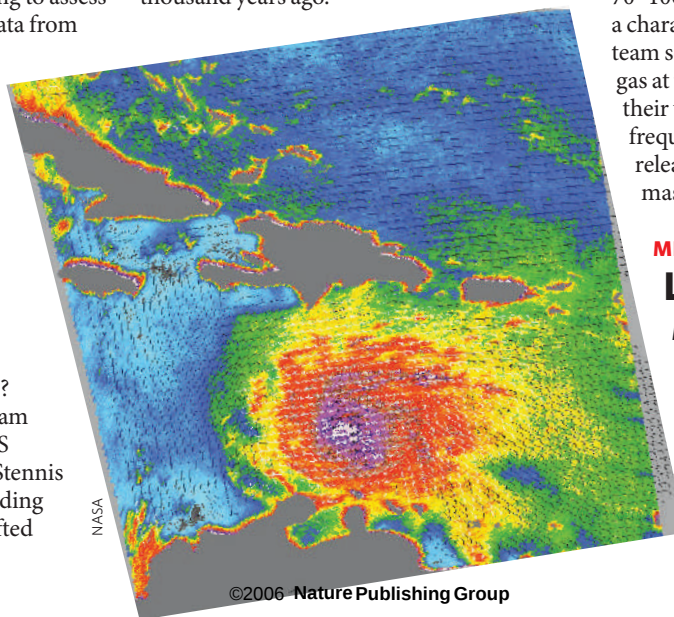
Researchers led by Nikos Poulakakis of the Natural History Museum of Crete collected and compared DNA up to 800,000 years old from a range of island and mainland elephants. Not all pygmies descended from the genus *Elephas*, as had previously been assumed. Those on Crete evolved from *Mammuthus*, making them more properly pygmy mammoths.



N. POULAKAKIS

around 100 million cubic metres of sediment as it stormed into the Gulf of Mexico in September 2004.

The researchers based their estimate on data from six instruments moored on the sea floor under Ivan's path. They suggest that the accumulation of sediment and shaking of the sea floor caused by a hurricane might trigger an underwater landslide, called a slump, resulting in a tsunami. Geological evidence suggests that just such a slump occurred in the Mississippi Delta several thousand years ago.



MEASUREMENT SCIENCE

Small scale

Nano Lett. **6**, 583–586 (2006)

Researchers in the United States have fine-tuned a sensing technique to measure mass with zeptogram (10^{-21} g) resolution. In principle, this should make it possible to weigh tens of atoms or a single molecule.

To make the measuring device, a team of scientists at the California Institute of Technology used silicon carbide beams just 70–100 nanometres in width that vibrated at a characteristic, very high frequency. The team squirted streams of xenon and nitrogen gas at the beams and recorded the change in their vibrational frequency. Comparing the frequency shift to the quantity of gas released yielded a reliable measurement of mass to the zeptogram scale.

METHODS

Lighting the way

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0509881103 (2006)

A new optical technique can concentrate and steer clusters of biomolecules such as proteins in solution.

Developed by Dean Hafeman of Protein Discovery in Knoxville,

Tennessee, and his colleagues, the technique uses a laser spot to concentrate charged biomolecules into an area measuring about 1 mm across. The laser is shone onto a photoconductive electrode, which sits below a conductive fluid layer containing the biomolecules, sandwiched by a second electrode.

The electric field created by the laser spot forms a trap which carries the molecules as the laser moves. The team demonstrates the process — which they call photo-electrophoretic localization and transport — by dragging a packet of proteins about half a millimetre per minute.

MOLECULAR BIOLOGY

Slide and jump

Nature Struct. Mol. Biol. doi:10.1038/nsmb1086 (2006)

The action of an enzyme that attenuates the infectivity of HIV has been shown to have an intriguing asymmetry.

APOBEC3G, which is encapsulated in HIV, helps stop the virus infecting certain cells. It does this by mutating the single-stranded DNA copied from the virus's RNA. Specifically, the enzyme converts a cytosine nucleotide from the end of a CCC triplet into a uracil nucleotide.

Myron Goodman and his colleagues at the University of Southern California in Los Angeles looked at why the mutations predominantly appear on the same end of the triplets. They show that APOBEC3G can slide and jump along single-stranded DNA in both directions to reach its target, but that it only catalyses mutations when moving in one direction.

IMMUNOLOGY

No pain, no heartache

J. Clin. Invest. doi:10.1172/JCI27540 (2006)

There may be a way to get the pain-relief benefits of the drugs called COX2 inhibitors without the cardiovascular problems that can result from long-term use.

Such drugs are thought to relieve pain and inflammation by suppressing COX2-derived prostacyclin (PGI2) and prostaglandin E2 (PGE2). Working in mice, Garret FitzGerald at the University of Pennsylvania School of Medicine, Philadelphia, and his colleagues show that disabling COX2 also causes blood-pressure and clotting problems.

They found that disabling a third protein in the pathway, microsomal PGE synthase-1 (mPGES-1), might be one way around this problem. It lowers the activity of PGE2 and still combats pain — but it boosts that of PGI2 and apparently dodges the cardiovascular problems.

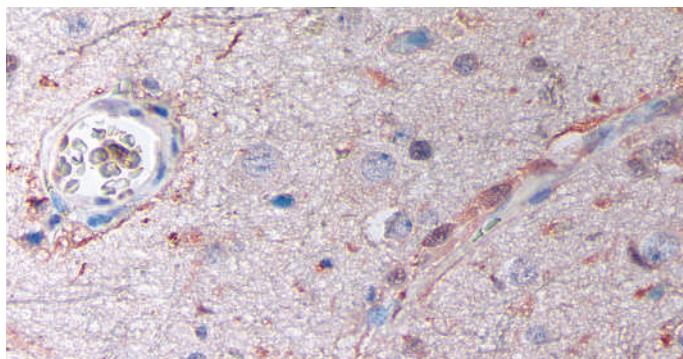
CANCER BIOLOGY

Stemming tumour growth

Cancer Cell **9**, 287–300 (2006)

Researchers hoping to stop the growth of gliomas, an aggressive form of brain cancer, have narrowed in on a promising target.

Howard Fine of the National Cancer Institute in Bethesda, Maryland, and his colleagues showed that a growth factor



ELSEVIER

associated with the cancer, stem-cell factor, aids the tumours' spread by promoting the formation of new blood vessels, a process known as angiogenesis. Drugs targeting stem-cell factor might slow this. The picture above shows tumour cells surrounding blood vessels.

MATERIALS SCIENCE

Sink or swim

Macromolecules doi:10.1021/ma060006u (2006)

Magnetic particles with collapsible polymer coats could be useful in catalysis, say researchers from the University of Düsseldorf in Germany.

Annette Schmidt and her colleagues made round magnetite particles that bristled with long-chain polymers and added them to methanol. Below a critical temperature, the particles precipitated, because their polymer chains crumpled close to the surface. Above it, the chains straightened out, helping the particles to disperse through the liquid. The researchers suggest this thermoresponsive system could be used to disperse catalyst-carrying particles during a reaction, then collect them up afterwards.

JOURNAL CLUB

Shahid Naeem

Columbia University, New York

An ecologist ponders the pendulum of natural history's paradigm.

The central paradigm of natural history has been stuck for 200 years, ever since Georges-Louis Leclerc de Buffon lost out to Carl Linnaeus.

Buffon championed an encyclopaedic approach to classifying nature, but his 36-volume attempt to study all aspects of all species, *Histoire Naturelle, Générale et Particulière*, often got fundamental things such as species relationships wrong.

In contrast, Linnaeus generally got species relationships right, by focusing on homologous, or structurally similar, traits. The homologous trait of feathers correctly flocks birds together, for example, whereas the functional trait of wings would erroneously group birds with bats and bees.

When Darwin later identified common descent as the rational basis for homology's success, it seemed the pendulum of natural history's paradigm was fixed.

But natural history based on homology has led to a biotic inventory that cannot predict the consequences of biodiversity loss, because it lacks information on species' functional traits. We need to know what species do, not who they are related to.

So when my graduate student pointed out a paper by Brian McGill of McGill University in Montreal and his colleagues (B. J. McGill *et al.* *Trends Ecol. Evol.* **21**, 178–185; 2006), I was fascinated to hear the echo of my lament among community ecologists.

McGill *et al.* claim that 50 years of concentrating on population dynamics has created a field that cannot predict species distributions. To fix this, they argue for an approach centred on functional traits.

This paper will, I'm sure, be widely read and debated. When the pendulum of a paradigm stops, someone needs to give it a nudge.

NEWS

Shifting constant could shake laws of nature

From the speed of light to the charge on an electron, the fundamental constants of physics had been assumed to be immutable. But that comfortable assumption is being challenged.

The latest in a series of experiments to question this view suggests that over the past 12 billion years, the ratio of the mass of a proton to that of an electron may have decreased. The result has left physicists curious, sceptical, and more than a little stumped.

Protons are about 1,836 times heavier than electrons. The exact mass ratio can be calculated by observing how a cloud of hydrogen molecules (each composed of two protons and two electrons) absorbs ultraviolet laser light. Wim Ubachs of the Free University in Amsterdam, the Netherlands, and his team have done just that, producing data hundreds of times more accurate than those obtained from previous experiments (see *Nature* doi:10.1038/news060417-7; 2006).

Using the Very Large Telescope in Paranal, Chile, they compared their results obtained with hydrogen in the lab with observations of light from two distant quasars. This light shines through clouds of hydrogen around 12 billion light years away. The lab result was smaller by 0.002% (E. Reinhold *et al. Phys. Rev. Lett.* **96**, 151101; 2006).

A change of 20 parts per million over 12 billion years isn't large — “not jelly”, as Andy

Fabian of the University of Cambridge, UK, puts it. But it could point to previously unknown subtleties in the way the Universe is put together. Such an effect is not explained by anything in physicists' standard model of particle physics.

The result has a confidence level of about 3.5-sigma, a statistical term that translates into a 0.3% possibility that it could be down to chance. That's good enough to be called an “indication”, says Ubachs, but for such an important potential result it is not a cast-iron observation.

“They’ve done the best job of anyone so far on comparing the proton and electron mass.”

“You don't want to book a ticket to Stockholm on a 3.5-sigma result,” chuckles John Webb, a physicist at the University of New South Wales in Sydney, Australia, who has also studied changes in the proton–electron mass ratio. “But they’ve done the best job of anyone so far on comparing the proton and electron mass.”

The mass-ratio effect has until now received less attention than the fine-structure constant, α , a measure of the electromagnetic force that keeps electrons in place inside atoms and molecules. Webb has been at the forefront of efforts to probe whether or not the mass ratio changes over time (J. K. Webb *et al. Phys. Rev. Lett.* **87**, 091301; 2001). He expects to publish his most detailed study later this year, which relies on light from many more quasars than used in previous analyses.

But the work is controversial. Webb says he has been criticized by senior astrophysicists



Weird stuff: observations of matter billions of light years away suggest the proton–electron mass ratio has changed over eons.

for even tackling the problem. “It’s as though you’re knocking a pillar of physics,” he says.

John Barrow, a cosmologist at the University of Cambridge, adds that astronomers are often more sceptical than physicists simply because they are more aware of how complicated —

NASA

US particle physics fights for survival

The United States must make a bid to host the International Linear Collider (ILC), according to a report out this week from the National Academy of Sciences. If the multi-billion-dollar particle accelerator is not built on US soil, the nation's high-energy physics community may be doomed.

High-energy physicists globally have endorsed the ILC as the next big project for the field. The collider,

30 kilometres in length, would smash together electrons and their anti-particles, known as positrons, at 500 GeV in the hope of probing fundamental particles, including the yet-to-be-discovered Higgs particle.

“Without a serious bid, the community is going to atrophy,” warns Harold Shapiro, an economist from Princeton University who chaired the panel that produced the report; the

panel was convened by the US Department of Energy and the National Science Foundation, which together fund the lion's share of high-energy physics research in the United States. In addition to recommending that the government make a strong bid for the project, the panel suggests that it invests in high-energy astrophysics and neutrino research. There are currently around 2,000

high-energy physicists working in the United States.

The report comes at a precarious time for high-energy physics in America. Earlier this year, budget cutbacks forced the Relativistic Heavy Ion Collider to take a private donation to continue operations, and the Tevatron, the nation's premier accelerator at the Fermi National Accelerator Laboratory, or Fermilab, in Batavia, Illinois, is expected to



and potentially error-prone — quasar spectroscopy can be. Fabian is certainly cautious about Ubachs' result. "Extraordinary claims require extraordinary evidence," he says, pointing out that many results at a similar confidence level turn out to be wrong. The most

likely error source lies in assumptions about the behaviour of the distant hydrogen cloud, he says. Some parts of the cloud could be hotter or moving faster than other parts, and the hydrogen might be mixed with a smattering of other elements.

Even if further studies do push the confidence level across the five-sigma threshold that physicists regard as convincing, the reason for the changed mass ratio is not understood, nor whether it is an ongoing effect. If true "the laws of physics as we currently understand them are incorrect at their very core," says Michael Murphy of the University of Cambridge, who works with Webb. "A new set of physics laws must be found which explain the new observations."

It is unlikely that protons are simply losing weight. But various versions of string theory suggest that extra dimensions occupied by a particle might affect properties such as its mass. Subtle changes in these dimensions could make physical constants vary slightly, explains Barrow. However, "there's absolutely no observational evidence to support this vast array of ideas," cautions Fabian. The paucity of hard evidence for string theory may be partly responsible for the upsurge in interest in variable constants, Barrow adds; results like Ubachs' could eventually provide a good way to assess the ideas. "I'm sure we'll see some theory papers about this," he says. "I might write one myself."

Fabian agrees that the problem has been receiving more attention over the past few years, but that "it's still a minority interest". The research needs intensive work on the very biggest telescopes, "a large investment in something that could turn out to be zero". But he agrees that it is an important problem to tackle: "Let's keep shaking the pillars to make sure they're rigid."

Mark Peplow

shut down within three to four years (see *Nature* 435, 728–729; 2005). "There doesn't seem to be a next step in place," Shapiro says. "This is a moment where important decisions have to be made."

Although the report will strengthen scientists' case for the project, Mike Lubell, head of public affairs for the American Physical Society in Washington DC, cautions that the collider —

estimated to cost at least \$6 billion — still faces an uphill battle. "In an era where the government is not keen on

raising taxes or cutting defence funding, a big project like this is going to run

into a lot of problems," he says. "I think this report will be very welcome in Europe," says Brian Foster, a physicist at the University of Oxford, UK, who is heading the European design

effort for the collider. Foster says that the Large Hadron Collider, a \$2.5-billion accelerator being built at CERN, the European particle-physics laboratory in Geneva, Switzerland, is consuming most of the continent's resources. A strong US bid for the ILC, he says, will strengthen the prospects of the collider, although he adds that Russia, Japan and China have all expressed interest in hosting it. ■ Geoff Brumfiel

ON THE RECORD

"The challenge will be to produce large quantities of this glue without it sticking to everything."

Indiana University bacteriologist Yves Brun discusses plans to mass-produce a powerful, natural glue formed by *Caulobacter crescentus*.

"A few years ago this kind of thing happened in Britain."

Chinese government spokesman Qin Gang rebuts a report from the British Transplantation Society that accuses China of harvesting organs from executed prisoners. Britain executed its last prisoner in 1955; Gang did not elaborate.

Sources: Indiana University, Reuters

SCORECARD



Cephalopod quirks
Squids are found to have genetically inherited personality traits, such as shyness.



Climateprediction.net
A programming glitch delays the distributed computing climate-modelling project by months.



Campus health
A mumps outbreak sweeps the US Midwest, prompting health officials to administer thousands of vaccines.



Medicinal pot
The US Food and Drug Administration declares there is no proof that smoking marijuana eases illness.

NUMBER CRUNCH

A study in *The Lancet* quantifies the US government's return on 28 phase III clinical trials funded by the National Institutes of Health.

\$335 million was the total cost of the trials.

470,339 years of healthy life will be gained by patients over a ten-year span.

\$15 billion is the estimated economic benefit that these trials will provide over the next ten years.

Source: Johnston, S. C. et al. *Lancet* 367, 1319–1327 (2006)

World Bank defends efforts to curb malaria

The World Bank has been forced to rebut a series of stinging criticisms from public-health commentators. They have accused it of failing to fulfil its promise to provide funds and leadership for Africa's fight against malaria.

Last year, the World Bank made a fresh pledge to implement global antimalaria schemes worth up to US\$1 billion over the next five years. This week's accusations mark the latest chapter in an acrimonious battle between the bank and a group of advocates who think it should hand the pledged funds over to agencies that, they argue, have greater public-health expertise.

The bank pledged in 2000 to make between \$US300 million and \$500 million available to African governments for antimalaria treatments and infrastructure. But the bank's critics point out that by 2005, no more than \$150 million had been provided.

"In 2000, they made the largest ever promise to spend money on malaria in Africa," says Amir Attaran, a law professor at the University of Ottawa, Canada, and lead author of a commentary that sets out the criticisms in this week's issue of *The Lancet* (A. Attaran *et al. Lancet* doi:10.1016/S0140-6736(06)68545-0; 2006). "It's now 2006, and they have broken that promise." The commentary is co-authored by 12 epidemiology and public-health experts, from Africa, Europe and North America.

Bank officials admit that their initial drive

was unsuccessful, partly because they did not receive as many applications from African governments as they expected. But they point out that the 2005 pledge, which aims to help the global Roll Back Malaria initiative hit its 2010 target of halving malaria cases, is already making progress.

"These criticisms are a year out of date," Suprotik Basu, a member of the bank's malaria team, told *Nature*. He says that, since the effort was relaunched last year, projects worth a total of \$190 million have been approved in Eritrea, Niger, Zambia and the Democratic Republic of Congo, and that the total will stand at \$400 million by July next year.

The bank will provide roughly half of the money itself. It will broker deals with other banks and with non-government organizations such as the Seattle-based Bill and Melinda Gates Foundation to raise the remainder of the cash. But Attaran complains that the bank, which has an annual tax-funded budget of \$20 billion, initially promised to raise all the funds by itself.

Charges levelled

More seriously, Attaran and his colleagues accuse the bank of massaging figures from previous malaria projects in Brazil and India to make them seem more successful. They also charge it with endorsing treatment options in India that will leave millions exposed to resistant forms of the disease.

"If you woke up with a fever, you wouldn't ring your bank for help."



Each year malaria kills more than a million people — mostly children in Africa.

Basu describes the allegations as "flat wrong". Data used to compile the bank's figures, given to Attaran upon request, came from peer-reviewed studies, and were neither collected nor doctored by the bank. And Basu insists that the bank is working closely with the World Health Organization (WHO) to ensure that the treatments it funds are recommended ones.

Whereas most malaria in Africa is spread by the *Plasmodium falciparum* parasite and ideally treated using drugs called artemisinins, around half of India's cases are caused by *P. vivax*, and can be treated using the much

C. BOURONCLE/AFR/GETTY IMAGES



Too late? The Chinese river dolphin may already have succumbed to pollution in the Yangtze River.

M. CARWARDINE/NATUREPL.COM

Last hope for river dolphins

The world's most critically endangered cetacean, the Chinese 'baiji' river dolphin, may finally have a chance of being saved from extinction. But it could be too late; researchers who carried out a nine-day pilot search for the dolphins last month didn't find a single one.

The freshwater baiji (*Lipotes vexillifer*) once thrived in their only habitat, the Yangtze River, which runs through central China. But fewer than 100 dolphins are

thought to be left in the river, which has become a busy, polluted highway. "If the giant panda is China's symbol of the destruction of forests, the baiji stands for polluted waters," says Wang Ding, from the Wuhan Institute of Hydrobiology.

An international team of scientists, led by Ding, is hoping to catch the animals and release them in a safer place, possibly the Shishou reserve, which is a 20-kilometre arm off the Yangtze.



cheaper drug chloroquine. Resistance to chloroquine is rife among *P. falciparum*, but Attaran says that, in areas of India where both forms are prevalent, patients are treated with the drug anyway, with partial success.

Basu insists that this is in accordance with WHO recommendations. But Attaran and his co-authors retort that the bank should stop trying to lead global efforts against malaria and instead turn its funds over to dedicated health agencies such as the Global Fund to Fight AIDS, Tuberculosis and Malaria. "If you woke up with a fever, you wouldn't ring your

bank for help," Attaran says. "The World Bank is great at being a bank, but bad at being the world's doctor."

In 2004, Attaran criticized the US Agency for International Development over the handling of its \$90-million malaria budget (see *Nature* 435, 257; 2005), and he has argued angrily with World Bank officials in the past. A bank spokesman accuses Attaran of "waging war against the bank". The spokesman claims that Attaran refused a 2002 invitation to discuss the issue with bank officials. ■

Michael Hopkin

But first the researchers — from China, the United States, Britain and Switzerland — need to find out exactly how many of the dolphins remain and where they are. They are preparing to conduct a search of 1,700 kilometres of the river in November, but carried out a pilot survey in March to refine their techniques.

The baiji are so few and far between that the best way to spot them is with acoustic devices. But that's a challenge. "The river is so noisy you

"If the giant panda is China's symbol of the destruction of forests, the baiji stands for polluted waters."

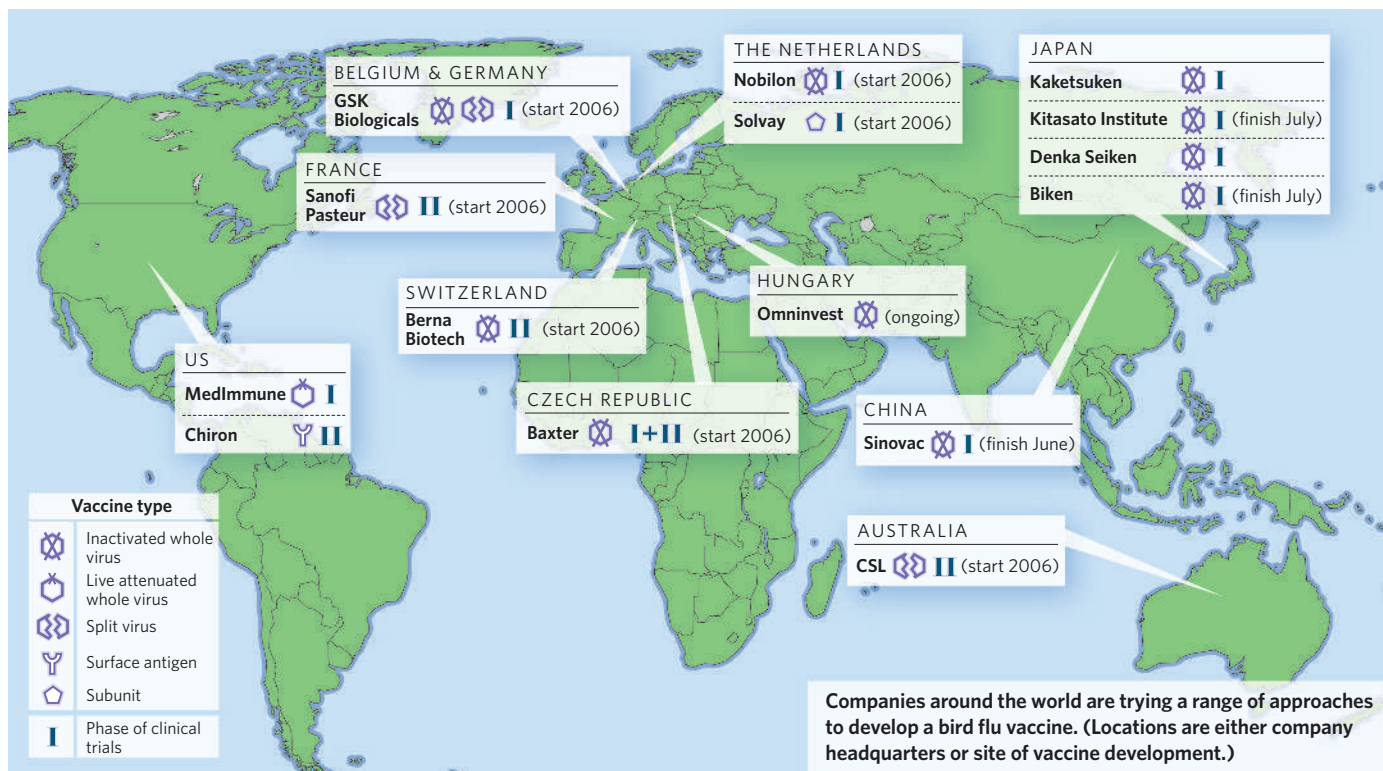
can't use traditional acoustic equipment," explains Jay Barlow, a marine mammalogist from the US National Marine Fisheries Service in La Jolla, California, who was on last month's cruise. He and his colleagues

are working on a method to clean up recordings from hydrophones, to isolate the baiji's distinctive whistles.

The researchers were disappointed not to see a single baiji on their recent search, but their hopes are now focused on the full-scale survey in November. "If none are found then, the burden of proof will change," says Barlow. "The species will be considered extinct unless proven otherwise." ■
Rex Dalton

**BIRD FLU IN FOCUS**

Find all of Nature's coverage on avian flu in one place.

www.nature.com/news/infocus/birdflu.html

PARTIALLY SOURCED BY IFPMA

Flu-vaccine makers toil to boost supply

The threat of a bird-flu pandemic is forcing vaccine developers to make tough choices about the types of vaccine to pursue. Many advanced trials are testing vaccines that use disrupted viruses, which have few side effects. But some experts now argue that, with such a wide gap between how much vaccine would be needed in a pandemic and how much the world's manufacturers could produce, attention should focus on vaccines that use the intact virus. Although thought to be riskier, they are far more effective at lower doses.

The likely shortfall in the available vaccine is highlighted this week by a study modelling the effectiveness of various interventions for mitigating a bird-flu pandemic. Using Britain and the United States as examples, Neil Ferguson at Imperial College London and his co-authors found that, to curb the spread of the disease, vaccinations would need to begin within one to two months of the pandemic starting (N. M. Ferguson *et al.* *Nature* doi:10.1038/nature04795; 2006). And a stockpile of vaccine for 20% of the population would greatly reduce the number of cases, even if it wasn't a perfect match for the pandemic strain.

"There is a significant gap between the desir-

able and actual access to the vaccine," says Klaus Stöhr, an adviser on flu-pandemic vaccine development at the World Health Organization. "There need to be fundamental changes in the way we produce and use influenza vaccines."

More than a dozen groups are developing pandemic vaccines, testing a range of strategies to boost potency and production capacity (see map). Some companies are testing more potent vaccine 'adjuvants,' which are used with a vaccine to boost the immune response. Others are developing cell-based systems to grow the virus, rather than the chicken eggs currently used, to increase vaccine yields.

But some experts argue that a quicker path to a greater vaccine supply is to switch to using whole (but inactivated) virus, rather than the 'split' virus conventionally used. The intact virus triggers a more powerful immune response, and such vaccines would potentially increase stocks, because a lower dose could be used.

"Companies should be using the whole virus, because there would be more to go round," says Graeme Laver, a virologist who has worked on flu for more than 40 years, primarily at the Australian National University in Canberra.

The risk is that such vaccines could trigger

toxic reactions — whole-virus flu vaccines used in the 1960s sometimes caused problems such as fever and pain, especially in children. So companies switched to split viruses, in which the viral coat is disrupted with a detergent.

But David Fedson, former medical director of Paris-based Sanofi-Aventis, believes the threat of bird flu makes whole-virus vaccines an attractive option again. "When death is the other option, a sore arm is not a bad price to pay," he says.

Several companies are in the early stages of developing whole-virus vaccines (see map). But many vaccine makers, including Sanofi Pasteur, the vaccine arm of Sanofi-Aventis, say they won't be switching to whole virus in the near future, mainly because their facilities are geared to making split-virus vaccines. "If a pandemic was declared this year, we would go with what we already have up our sleeve," says Rachel David, a spokeswoman for Melbourne-based CSL, which has a split-virus vaccine in clinical trials.

But Stöhr cautions against being too obsessed with the whole-versus-split virus issue. "There are many avenues for improving the vaccines," he points out. He says a global action plan is urgently needed to prioritize research efforts and rally more funding to develop flu-pandemic vaccines.

Carina Dennis



Feel it in your bones

Deciding whether our ancestors evolved as a single lineage may depend more on philosophy than fossils. **Rex Dalton** reports.

The human family tree may now be a lot cleaner, according to the discoverers of a cache of fossils on a hillside in Ethiopia.

The researchers say the fossils, from the Afar region, go a long way towards showing that early hominids were not a diverse group of coexisting species. The team suggests three species evolved as a single lineage between at least 4.4 million years ago and 2.9 million years ago — an era when humankind refined its ability to walk upright while developing new ways to live (see timeline).

The idea is one of the most contentious in palaeoanthropology. The fossil trove, reported earlier this month (T. D. White *et al.* *Nature* **440**, 883–889; 2006) has confirmed some important aspects of the trail towards the genus *Homo*, which appeared around 2.3 million years ago. But experts are still bickering over the relationship between the species that have been found.

The latest report describes 30 fossils, belonging to at least eight individuals of *Australopithecus anamensis*; before the discovery, the only known specimens of this species came from northern Kenya, about 1,000 kilometres to the southwest (M. G. Leakey *et al.* *Nature* **376**, 565–571; 1995). The Afar specimens date to 4.1 million years ago, and were dug from ravines cut by rain at a site called Asa Issie, meaning red hill. They are the latest discovery made by the Middle Awash project, an international team that has explored these geological formations for nearly 25 winters.

An hour's walk to the northeast of Asa Issie

is the Aramis site, where the team found fossils of an earlier species called *Ardipithecus ramidus*, dated to 4.4 million years ago. An hour's walk to the east lie fossils of *Australopithecus afarensis* — the species to which Lucy belongs, dated to 3.4 million years ago. Many consider *A. afarensis* to be the immediate predecessor of the genus *Homo*. It was first identified at Hadar, about 70 kilometres north of Asa Issie, and lived until at least 2.9 million years ago in eastern Africa.

Theory of relativity

An understanding of our own species' lineage is crucial to building a picture of how early hominids evolved, their gait, diet and tools, and the effects of the environment. Experts have squabbled over the relationship between *Ar. ramidus*, *Au. anamensis* and *Au. afarensis* ever since they were discovered: some believe each species evolved directly into the next, and others insist that they lie on separate branches of the family tree.

The latest Afar discovery is exciting experts because it shows that the three hominids existing in the same area, but in successive time periods. Tim White of the University of California, Berkeley, co-leader of the Awash team, believes this points to a direct lineage between the three — a process called phyletic evolution.

The new *Au. anamensis* fossils are only 300,000 years younger than *Ar. ramidus*, meaning that if one became the other, the

changes would have had to happen that fast. But the key point, says White, is that fossils of *Au. anamensis* and *Au. afarensis* have never been found in sediments the same age as those containing *Ar. ramidus*. If fossils of the different species were found together, that could show that they belonged to multiple lineages existing simultaneously.

Finding remains of all three species in the same area but not from the same time period suggests they did not coexist, says White. The team confirmed the dates using several geological dating methods on layers of rock above and below the fossils, along with examinations of other animal fossils found, including antelopes, monkeys and pigs.

The specimens also provide anatomical clues to evolutionary history. "The new *Au. anamensis* fossils are anatomically intermediate between the earlier *Ar. ramidus* and the later *Au. afarensis*," says White. For example, the teeth of the newly discovered *Au. anamensis* fossils seem adapted to chew tougher and more abrasive foods than *Ar. ramidus*. The researchers believe this shows that *Au. anamensis* had a broader diet.

"All this strengthens the view that there is phyletic evolution from *Ar. ramidus* through *Au. anamensis*," says White. He believes he has nailed down the relationship between the two later species,

although he says that further specimens are needed to prove the earlier link.

This month's *Nature* paper makes a bold argument, and shows the Awash team seeking to put its mark on the record. Others in the field are impressed. "When you find 30 new hominid fossils, you are allowed a certain amount of conjecture," says Bernard Wood, a palaeoanthropologist at George Washington University in Washington DC. "As always, they have done a fantastic job."

"When you find 30 new hominid fossils, you are allowed certain conjectures."



Rock on: members of the Middle Awash project have been hunting hominid fossils in Ethiopia for 25 years.

R. DALTON

But he and others are unconvinced by the Awash team's conclusion: "This is only the first half of the rugby match," says Wood.

Meave Leakey, lead author on the *Au. anamensis* discoveries in Kenya, is more blunt. "I don't believe this," she says. "We do not have the specimens to fill the gaps."

Ancestral pile-up

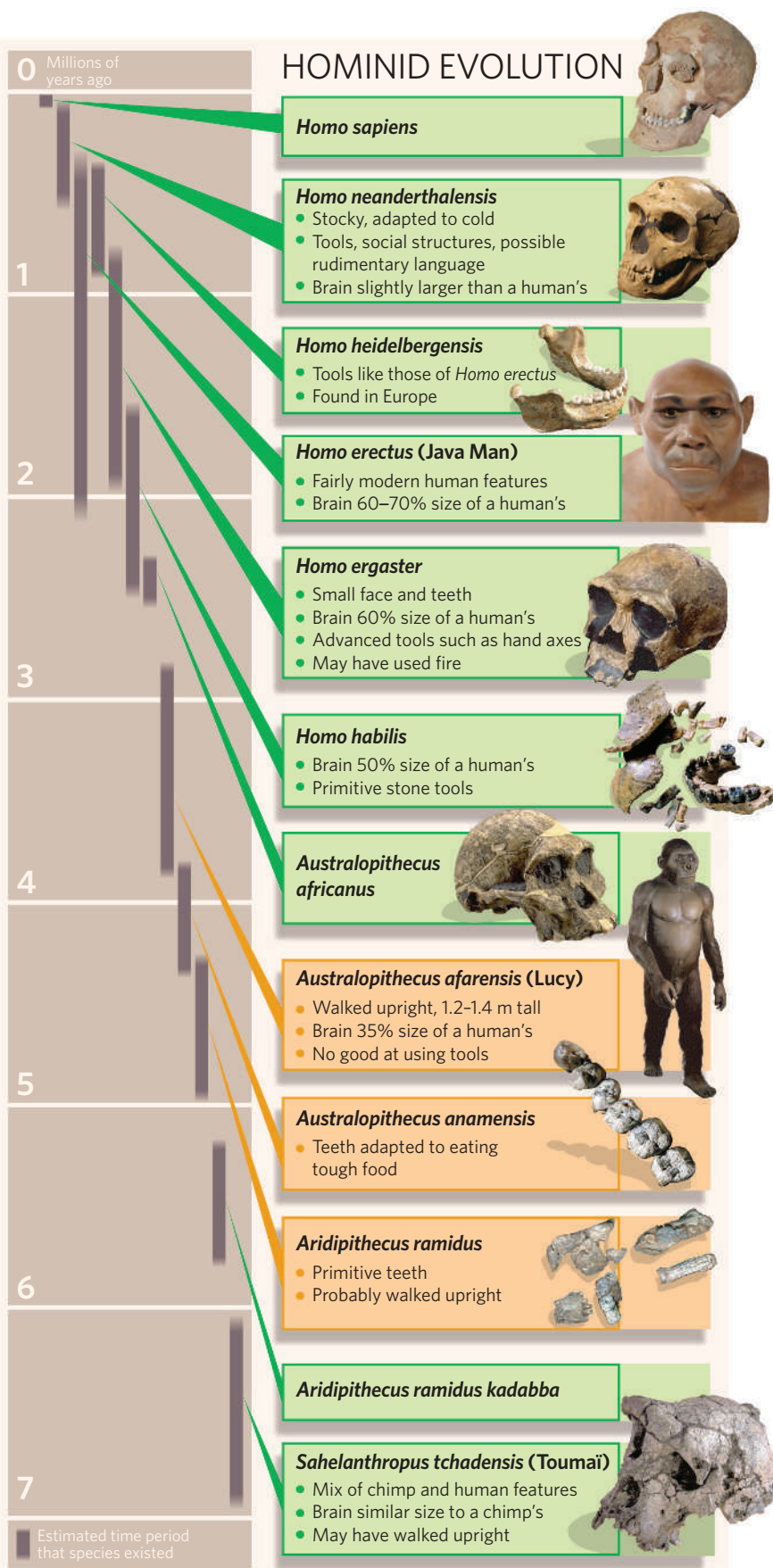
Leakey and Wood are among those who believe that other, as yet undiscovered hominid species may have lived at this time, from 4.4 million to 2.9 million years ago. The existence of other species would cloud or eliminate the argument for a direct lineage. "My prejudice is there are more lineages rather than fewer — more diversity," says Wood. "I have to concede these new data are dramatic. But we should beware coming out with a complete explanation when we don't have all the evidence."

This argument frustrates White. "There were Martians there back then too," he says. "And spacecraft all over the Pliocene — we just haven't found them yet."

Similar arguments run for various phases of hominid evolution, for example whether *Homo ergaster* evolved into *H. erectus*, or whether they were two coexisting lineages — White advocates the former. But ultimately, the argument comes down to the point that more fossils could always be found, so it is unclear that the two sides will ever agree.

In the meantime, the new *Au. anamensis* fossils are at the Ethiopian National Museum in Addis Ababa, where they will augment the Middle Awash hominid collection, which spans nearly 6 million years. "The discovery at Asa Issie is really wonderful," says Mamitu Yilma, the museum's director. "This strengthens the museum as an ideal centre to study human evolution."

TOP TO BOTTOM: P. GOETGHELUCK/SPL; O. SCHÖTENSACK; NAT. HIST. MUS.; P. GOETGHELUCK/SPL; J. READER/SPL; SMITHSONIAN; J. TRUEBA/MSF/SPL; T. WHITE; S. SIMPSON; M. BRUNET



Korean women launch lawsuit over egg donation

Two women who donated eggs to the South Korean effort to produce cloned human embryos for research have filed a lawsuit against the state and the two hospitals that collected the eggs.

The women filed the suit on 21 April claiming that they were not made aware of the potential risks of donation. They are seeking 32 million won (\$35,000) in compensation.

Woo Suk Hwang of Seoul National University used more than 2,200 eggs from 119 donors in his failed attempts to clone human embryos. The National Bioethics Committee reported in February that there were serious breaches in the informed-consent procedures used to collect the eggs. Many women were not told of the risks and a large number have since reported side effects.

California court quashes action over stem-cell plan

In a much-anticipated decision, a state judge has ruled that California's \$3-billion stem-cell research programme is legal,

bringing the state closer to major funding of such studies.

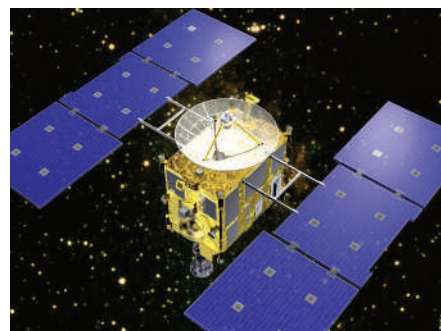
The ruling on 21 April held that the initiative overwhelmingly approved by voters in 2004 is constitutional and is being appropriately administered by the California Institute for Regenerative Medicine (CIRM). A loose-knit group of religious and anti-tax organizations sued the institute to block the research programme (see *Nature* 437, 800–801; 2005) — the group now says it expects to appeal. If it does, a final decision by the state supreme court could take a year or more.

CIRM officials say that last week's ruling should allow the state to sell up to \$200 million in bond-anticipation notes for research. Full bond funding would follow if the decision is upheld.

Japan lays plans for re-run of Hayabusa mission

Japan's space agency has created a 20-member team to prepare for new missions to an asteroid and to the Moon.

By 2010, the Japan Aerospace Exploration Agency (JAXA) aims to launch a successor to the Hayabusa probe. That probe visited the asteroid Itokawa in November



JAXA

Japan's next attempt to collect dust from an asteroid will use a probe very similar to Hayabusa.

2005, but may not have collected rock samples as planned. Its successor will go for a clear sample return from a different asteroid, says Jun'ichiro Kawaguchi, the mission's supervisor. He says the probe will be a rebuild of the Hayabusa design, with improvements to the device used for sample collection and better chemical engines to avoid the problems that have put Hayabusa's safe return to Earth in doubt.

A second project aims to land a different probe on the Moon by 2015. It would follow next year's scheduled launch of the agency's Selene lunar orbiter.

The new push is part of JAXA's attempts

to realize its 20-year vision released last year, which was widely criticized due to the lack of concrete ideas (see *Nature* 439, 132–133; 2006).

Deal ends dispute over Harvard's missing samples

A couple accused of stealing research material from Harvard University six years ago has reached an agreement settling a federal criminal indictment against them in Boston, Massachusetts.

Researchers Jiangyu 'Jonathan' Zhu and his wife Kayoko Kimbara, both now living in San Diego, will have the indictment against them dropped in one year under terms of the 13 April agreement.

In the deal, they acknowledged that they improperly took 20 boxes, including 1,300 genetic samples, late at night when moving to a Texas university. They were arrested in 2002 in San Diego, where they had subsequently gone to conduct research (see *Nature* 417, 886; 2002).

The couple had been under electronic monitoring and home custody between 2002 and 2005 for interstate transportation of stolen property. All the material in question has since been returned.

Gold stars for work on little chemical sensors

In Texas, the Lone Star state, where everything is bigger, scientists have turned convention on its head by creating these miniature 'nanostars' (pictured).

Jason Hafner of Rice University in Houston and colleagues synthesized these tiny structures, which are roughly 100 nanometres across and made of gold. Each of the stars' tips has a unique spectral signature, and taken together they can be used to determine how the nanostar is oriented in three dimensions. As such, they could be used as chemical sensors, the scientists write (C. L. Nehl *et al.* *Nano Lett.* **6**, 683–688; 2006).



J. HAFNER/RICE UNIV.

Tissue donors don't own their cells, says judge

A US federal judge has issued a ruling that could have implications for deciding who owns biological samples. On 14 April, the judge ruled that Washington University in St Louis, Missouri, owns the tissue samples donated by patients for medical research.

The decision means that the university can once again make its tissue bank available to researchers. The repository holds samples from more than 30,000 men,

but was closed when litigation began.

The action was brought by surgeon William Catalona, who used the tissue bank's samples to develop a test for detecting prostate cancer. He says he plans to appeal against the decision.

Catalona sought control of the specimens in 2003, when he moved to Northwestern University in Chicago. When Washington University refused, Catalona claims that he got 6,000 donors to request that their samples be transferred to him. Washington University then asked the courts to decide who controlled the samples.

DIAL 'E' FOR ETHICS

Facing a moral dilemma in the lab? No reason to panic.

Helen Pilcher meets the academic troubleshooters who promise a quick answer to any ethical problem.

Sometimes you're midway into a research project before an ethical dilemma reveals itself. This was Joachim Hallmayer's experience at the Stanford School of Medicine in California. About a year after he and his team had begun recruiting children for a genetic study of autism, they realized that they couldn't agree what results to share with the parents. "It's sometimes difficult to know what information will be useful and what will be dangerous," admits Hallmayer, a psychiatric geneticist.

But the team was in luck. Across campus, at the Stanford Center for Biomedical Ethics, a pilot project offering biomedical researchers speedy practical advice on ethical concerns was under way. Hallmayer contacted the 'dial-an-ethicist' project, or bench-side consultation service as it is also known, and three weeks later got professional answers to his questions.

Hallmayer's team is now pursuing its autism study with a renewed sense of harmony. But the Stanford service aims to go beyond ethical troubleshooting, say co-founders David Magnus and Mildred Cho. "A lot of scientists don't really see ethics as a part of their job," says Cho. By making the usually academic field of bioethics more accessible, Magnus and Cho hope to promote a culture of ethical thinking within the laboratory.

Often scientists don't think about ethics until it is too late — sometimes when their research has already hit the headlines. In South Korea, disgraced stem-cell pioneer Woo Suk Hwang not only fabricated results, he also obtained eggs from women donors under questionable circumstances. And he did it all while claiming to follow ethical guidelines. It's an extreme example of the harm that can be caused when ethics advice is ignored, but one



PHOTOTAKE/ALAMY

Stanford biomedical researchers can now phone a friend to discuss their pangs of conscience.

that raises questions about the role of bioethicists in the laboratory.

Some people doubt whether practical ethics advice can make a real difference. Can bioethicists retain credibility when their advice is sought but ignored? Are they liable when things go wrong? And what about claims that they rubber-stamp most research proposals?

"Bench-side consultations are a way of integrating ethical thinking into a scientist's everyday life," says Magnus, director of the Stanford centre. Like most of his colleagues he doesn't think bioethicists should be expected to prevent misconduct. But he believes his bench-side service can foster integrity in trainee scientists and so indirectly prevent research going off the ethical rails.

The Stanford service is designed to help researchers identify the ethical and social issues that arise in their work and aims to complement, not replace, the bodies that regulate human and animal studies. Institutional review boards (IRBs), for example, oversee all federally funded US biomedical human studies. They evaluate the risks and benefits to people who participate, from their recruitment through to the confidentiality of results.

An IRB seal of approval must be in place before a study begins.

Unlike IRB approval, the Stanford service is voluntary, not mandatory, and it yields confidential advice, rather than edict. The Stanford team will advise researchers at any point in a study, although they prefer to be involved at the start. The pilot seeks to address issues, such as the broader societal implications of a study, that go beyond the scope of IRBs. In fact, the project was set up, in part, to offer advice on human embryonic stem-cell research, which initially fell outside the IRB's purview.

Brisk business

Over the past six months, the pilot service has given consultations to seven different Stanford research groups. Topics ranged from oncology trials to microarray analysis, and the ethical issues from conflicts of interest to what to do with incidental findings. Six of the queries were easily resolved, most within 24 hours, and half of the responses involved alerting researchers to existing rules rather than developing new policies.

Hallmayer's request, however, required deeper analysis. The California Autism Twin Study (CATS) plans to assess 300 twin pairs on a range of skills, including intelligence, language and planning ability. To date, more

"It's sometimes difficult to know what information will be dangerous."
Joachim Hallmayer

than 60 pairs of children have been recruited and around a dozen tested. But as the data trickled in, ethical concerns grew about how much to tell the parents about the test results.

This is partly because the CATS tests can differ from those used by psychiatrists to diagnose autism and are administered by trained researchers rather than clinicians. Such non-clinical test results raise several problems. A presumed autistic child could, for example, test normal, or a normal sibling could display mild, sub-clinical problems. DNA tests designed to check whether twins are identical or not, and to exclude children with the inherited mental impairment, fragile X syndrome, create their own problems. A positive fragile X result would suggest one parent is carrying the causal mutation, for example.

Should we tell them?

Any such findings could have repercussions for family dynamics, insurance and state benefits, and so Hallmayer felt it wise to tell parents less rather than more. "They may have a right to know," he says, "but the results can give a child a label that is unnecessary and meaningless." Others on the team, however, felt it was wrong to withhold findings from parents.

Hallmayer contacted the Stanford ethics service last year. He and his team discussed their problems with four consultants, including Cho and Magnus. Three weeks later, they received a written list of recommendations.

The ethics team concluded that because the test results do not indicate a treatable or preventable life-threatening condition, researchers are not required to inform the parents.

Non-clinical test results should not be disclosed, they advised, but other results can be shared if the parents want to know. Positive fragile X tests should always be replicated in a clinical laboratory, and parents should be made aware of the possible implications for insurance and employment.

The CATS team tweaked its study design, received additional IRB approval, and pursued its study. Both Hallmayer and the Stanford ethicists are pleased with the outcome. But the bench-side experience is not always a happy one.

Bioethicist Insoo Hyun, from Case Western Reserve University in Cleveland, Ohio, spent three months in summer 2005 in Hwang's laboratory studying the cultural, social and legal implications of the Korean stem-cell research. He also helped to fine-tune Hwang's informed-consent procedure for egg donation, publishing a paper reporting the procedure in *The American Journal of Bioethics* (K. W. Jung and I. Hyun *Am. J. Bioeth.* 6, W19–W22; 2006). When he learned in December that Hwang had not followed the procedure he felt angry,



Parents may be better off not knowing the findings of studies on their autistic children.

saddened and betrayed.

Hwang seems to have obtained independent advice from at least two ethicists in addition to IRB approval for his research. But he violated ethical guidelines by collecting eggs from laboratory juniors and other donors who were not warned of possible side effects.

About 20 of the donors have developed side effects including abdominal pain, diarrhoea and vomiting. And their eggs contributed to research that now appears to have been faked. The kindness of the donors has been exploited, patient hopes shattered, and public confidence in stem-cell research is in tatters.

The scandal underscores the perils of ignoring ethical advice. But it also highlights the limitations of ethical consultations in the laboratory. Hyun and his co-author had to retract their paper when it became clear that Hwang had ignored their recommendations (G. McGee, *Am. J. Bioeth.* 6, W33; 2006). "My experience with Hwang has changed my attitude to getting involved with other labs," says Hyun.

In future, Hyun says he would be more sceptical. Magnus, who has also written about the ethics of egg donation, believes that the ethicists in Hwang's lab did little to find ethical flaws. "I am concerned that too many of the bioethicists there were cheerleaders," he says.

In his defence, Hyun says, bioethicists view themselves as pilots through treacherous waters. "Being an ethical consultant doesn't put you in a position to prevent misconduct," he says. Consultants presuppose that researchers have the integrity to take their advice seriously. "If people are out to deliberately deceive you, there's nothing you can do."

Some view all ethical oversight with suspicion. Bioethicists merely provide the veneer of a job well done, says Vera Sharav, president of

the Alliance for Human Research Protection, New York, a watchdog organization that seeks to expose unethical research practices. Most, she believes, issue ethical approval too easily and fail to offer truly independent advice.

But Ronald Green, a bioethicist from the Dartmouth College Ethics Institute, Hanover, New Hampshire, finds generalized accusations of 'bioethics for hire' preposterous. Green chairs the ethics advisory board at Advanced Cell Technology, a US-based biotechnology company developing stem-cell therapies. Any payment to ethicists is a minimal daily stipend, and with ethical advice carrying no weight in law, 'buying it' simply doesn't make sense, he says. For his part, Hyun believes his relationship with Hwang was sufficiently independent, with funding from a Fulbright scholarship, to allay such concerns.

Blowing the whistle

What would the Stanford ethicists do if they discovered research that was illegal or breached public health and safety? Although they sign confidentiality agreements with Stanford researchers, Magnus and Cho say they have an obligation to report misconduct if they come across it. But what about less serious transgressions? "I don't think bioethicists should be policing research," says Magnus.

To avoid potential criticisms, the Stanford team is financially independent thanks to federal funding. The service is now being offered to all biomedical researchers at Stanford; other institutions, including Case Western Reserve University and Duke University in North Carolina, are setting up similar bench-side consultations. Hyun is not involved in the Case Western service at present, although he has not ruled that out. The hope is that such a service will promote ethical awareness in the lab. It's not a watchdog, but it must avoid looking like a show dog. It might not prevent deliberate misconduct, but it should help researchers who want to do the right thing.

Helen Pilcher is a freelancer based in Nottinghamshire, UK.



"If people are out to deliberately deceive you, there's nothing you can do."
— Insoo Hyun

DIGITAL DIGS

Archaeologists are bringing past worlds vividly to life on the computer screen. But are the high-tech graphics helping science, or are they just pretty pictures? **Michael Bawaya** takes a look.

There's more than one way to sink a ship, as Donald Sanders knows. President of the Institute for the Visualization of History in Williamstown, Massachusetts, Sanders spends a lot of his time repeatedly sinking a vessel off the coast of Cyprus.

The ship isn't real — it's a computer model of a vessel that sank in the fourth century BC. Sanders is trying to recreate what happened when the ship went down, leaving nearly 500 intact amphorae, or storage vessels, to be found centuries later on the sea floor. By loading his ship with a virtual crew and cargo, then sinking it in a number of different potential disasters, Sanders hopes to find a sequence of events that closely matches the archaeological evidence, and so work out what might have happened centuries ago.

His project, which should be completed later this year, is just one example in the growing field of virtual archaeology. It is a highly visual way to recreate lost worlds, helping researchers to imagine environments long past. Since the field emerged in the early 1990s, archaeologists have created a number of virtual sites, including Pompeii and the tomb of Nefertiti.

Virtual archaeology can recreate individual objects, such as a hand axe, or entire sites and landscapes, such as Stonehenge — an ambitious project that took six months to complete. The possibilities of the technology are enormous, says Jeff Clark, director of the Archaeology Technologies Laboratory at North Dakota State University in Fargo. "Virtual archaeology can allow us to explore ideas in ways that are otherwise not possible," he says.

Sometimes, it may seem a bit too real. The technology is so impressive that it can eclipse the importance of the models, as users navigate their way through elaborately designed palaces or villages. Sarah Kenderdine, an archaeologist and curator at Museum Victoria in Melbourne, Australia, says some academics regard virtual heritage work as "kitsch and consumerist" rather than scientific and educational.

But even if the models do resemble a video game, they can also offer an unparalleled entry point into a vanished world, providing new and unexpected insight. For instance, a three-dimensional representation of Rome's Colosseum, created several years ago by a team at the University of California, Los Angeles, has shown researchers just how cramped some of the upper hallways really were — confounding their expectations.

Key to the new insights are technologies that let archaeologists peer into obscured artefacts

and document landscapes in great detail. Computed-tomography scans of Egyptian mummies, for instance, can reveal internal structures noninvasively. Laser and computed-tomography scans of rare and fragile human remains or artefacts produce three-dimensional models that can be measured and analysed in a way that real objects cannot. Some archaeologists are even taking the technology into the field, armed with laser scanners to create three-dimensional replicas of their findings as the excavation progresses.

Such data are then fed into a modelling or drawing program, and displayed on the Internet. Researchers who live thousands of miles away, or theoretically years in the future, can then access these items.

Looking back

It sounds expensive, but it doesn't have to be, says Robert Stone, director of the Human Interface Technologies Team at the University of Birmingham, UK. Early virtual reconstructions were powered by supercomputers that cost hundreds of thousands of dollars. Today, he says, experts can achieve the same effects, and more, with a standard desktop computer. But software remains expensive; Sanders, for instance, spent nearly US\$30,000 on the programs he uses to sink his virtual ship.

The time spent working on such projects can also increase the cost. At Sanders' Massachusetts laboratory, projects run the gamut from individual objects, such as a small model of an ancient Egyptian ship that might take a



In the drink: a computer's view of the lowest level of a Kyrenian ship that sank in the fourth century BC off the coast of Cyprus with a cargo of amphorae and millstones.



Lost glory: a recreation of the throne room of the Assyrian king Ashur-nasir-pal II, as it might have looked in the ninth century BC.

LEARNING SITES



VIZIN/KYRENIA SHIP PROJECT

day to complete, to entire building complexes, such as a ninth-century-BC Assyrian palace with hundreds of rooms, which could take a year to finish. The former might cost a few thousand dollars, whereas the palace could run into hundreds of thousands. "Doing virtual reality is becoming easier," he says, "but it still requires computer-modelling skills, data-interpretation skills and some programming to get it right."

In many cases, it also takes an interdisciplinary team that includes archaeologists, architects, art historians, programmers and graphic modellers. And they must be trained to understand the cultural importance of their work, says Maurizio Forte, an archaeologist at the Institute of Technologies Applied to Cultural Heritage in Rome, Italy. "Technology is only a tool," he says. "The method is the core."

Despite its capabilities, virtual archaeology has been slow to catch on. Sanders estimates that currently 80–100 projects are under way at any one time. Western Europe is the most generous funder of such research. Kenderdine, in Australia, taps into European sources to fund worldwide projects such as her recreation of the ancient Cambodian city of Angkor.

At the museum in Melbourne, visitors can experience Kenderdine's work inside a 'virtual room', an octagonal laboratory with panoramas displayed on the walls. The images aren't computer-generated, but are stereographic, high-resolution photographs. The idea is for visitors to be immersed in the ruins of Angkor, as if they were actually there.

Such detailed recreation requires considerable research. A virtual model must be geometrically complete, says Elizabeth Riorden, a

researcher at the University of Cincinnati, Ohio, who works on virtual reconstructions of the site believed to be the ancient city of Troy. "But we often have very incomplete knowledge of an archaeological site," she says. So researchers may be forced to resort to educated guesswork.

Some archaeologists are sceptical because they don't understand the methods and assumptions used to create models, says Sanders. And there are currently no generally accepted guidelines. A February 2006 conference at King's College London aimed to address the issue of transparency. For the first time, virtual-heritage experts agreed to draft specific guidelines for documenting the creation of virtual models. These might, for instance, tell the viewer that 75% of a model was built using data from the archaeological record, and the rest was based on assumptions.

Image consultants

However it is created, a virtual model may end up looking so realistic that, as Clark notes, "it imparts a sense of truth that in fact does not exist". Some virtual models could be partly accurate and partly not because one portion is based on abundant archaeological information whereas another portion is based on conjecture. "If you really try to get the visual simulation as accurate as possible, you must research the nitty-gritty details," he says.

Most researchers agree that the technology is making archaeology more accessible to the public, as models are featured in museums and television documentaries. Displays can be posted on the Internet, in visitors' centres and in schools. Archaeologists may in fact be the hardest to bring around. "Multidisciplinary communities embrace virtual archaeology as a key part of their research," says Forte. "Traditional archaeologists are still very conservative and think virtual archaeology is a game."

The cost and complexity of the work may keep some researchers away. One of virtual archaeology's grand ambitions, says Sanders, is "to recreate the past in all its complexity, with people, animals, climate, vegetation and the ability of the scene to change and adapt to external conditions". And the more detail needed, the greater the technical and financial demands (see 'Recreating ancient life'). "George Lucas could probably do it if he wanted to," he says.

But the evolution of video-game technology is helping virtual archaeology develop, says Stone. He recently gave his undergraduates the task of creating three-dimensional virtual models, complete with animation. Although none of the students is an experienced programmer, they used gaming software to make impressive models — ranging from an Afghan village with blowing sand and swaying trees, to an underwater wreck replete with marine life. It may be that virtual archaeology won't need someone like George Lucas after all. ■ Michael Bawaya is a freelance writer in Albuquerque, New Mexico.

RECREATING ANCIENT LIFE

Reconstructing virtual archaeological sites and landscapes is challenging, but simulating ancient life is even more so. Maurizio Forte, of the Institute of Technologies Applied to Cultural Heritage in Rome, Italy, has been working to incorporate artificial life, or A-life, components into simulations of past environments.

A-life simulations have been used in many fields, from predator-prey interactions in ecology to the rise and fall of

the Anasazi civilization in the American southwest. But Forte thinks it can add a much-needed dose of realism to virtual models of particular archaeological sites.

"With computer simulation," he says, "it becomes possible to build artificial societies of computational agents and perform experiments under laboratory conditions, trying out different configurations and observing the consequences."

Forte is testing the notion of A-life in his recreation of the

Via Flaminia, one of ancient Rome's most important roads, by incorporating it into reconstructions of a famous battle that occurred along it. The technology allows him to set the two armies against each other, running through all the possible sequences and unpredictable details that may have occurred during the battle. Historical records tell us the winner of the battle, but modern-day computers may tell us what probably happened.

M.B.

Magical mantle tour

Geophysicists are racing to understand a recently discovered phenomenon deep in the Earth. **David Cyranoski** joins them.

It's a place where everything goes weird. Some 2,700 kilometres inside the Earth, just above the boundary between the planet's solid mantle and its liquid outer core, is a layer where seismic waves act strangely. Instead of carrying on through the Earth, some waves speed up when they enter the zone and, once inside, can vary in speed depending on their direction of travel. In some pockets, they can slow by as much as 30%.

Such seismic waves are scientists' primary means of probing Earth's depths; waves bouncing through the planet's interior have revealed how Earth is layered into crust, mantle, outer core and inner core. But what could account for the seismic oddness of this region, known as the D" (or D-double-prime) layer?

Until recently, nobody even knew what questions to ask. "A couple of years ago, deep-Earth seismology seemed stagnant," says Christine Thomas, a seismologist at the University of Liverpool. But in 2004, a group led by Kei Hirose at the Tokyo Institute of Technology reported that they had reproduced the high pressures and temperatures of the lowermost mantle in the laboratory (see "A collaboration on the rocks"). And they found that the predominant mineral there — a magnesium silicate, MgSiO_3 — can take

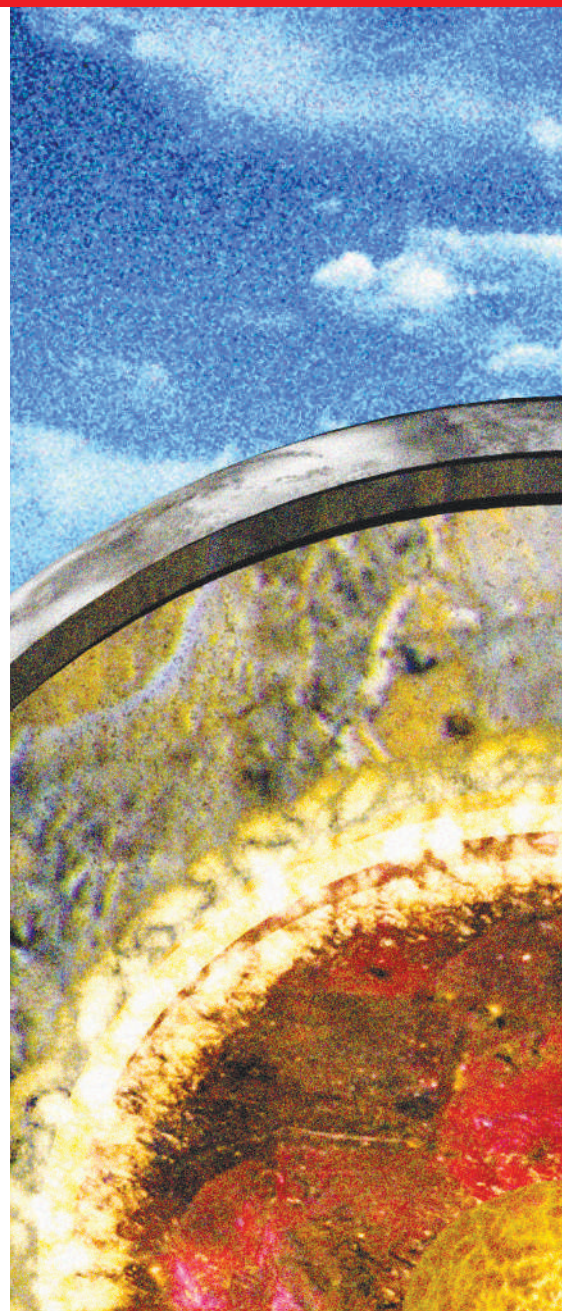
on a previously unknown structure¹.

The structure, known as the 'post-perovskite phase' of MgSiO_3 , has the same mineralogical composition as the more common perovskite form, but its atoms line up differently. "The denser post-perovskite is like a stack of two-dimensional sheets, compared with the three-dimensional bonding structure of perovskite," says Hirose (see graphic). And that shift in atomic arrangement, known as a phase transition, could help explain the seismic oddities of Earth's lower mantle.

Is that it?

Phase transitions are common in geology, and to some Hirose's explanation seems disappointingly mundane. "What we were searching for all those years was nothing more than a phase transition," says John Brodholt, a mineral physicist at University College London. Nevertheless, post-perovskite has energized seismology, geodynamics and mineral physics.

In a flurry of publications, researchers have debated the structure's properties, what other seismological mysteries it might explain, and its implications for heat flow within the mantle. Planetary experts are re-examining some quirks of Earth's rotation in light of the post-perovskite discovery, and are even thinking about how it could explain features of other



A COLLABORATION ON THE ROCKS

Post-perovskite MgSiO_3 was a long-anticipated discovery. How did it finally happen? That depends on whom you ask.

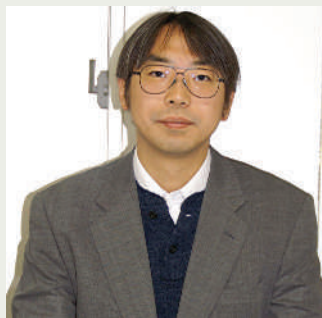
Kei Hirose and Shigeaki Ono used to be collaborators. Now, aside from polite greetings in the hallways, they don't talk.

They work one floor apart at Tokyo Institute of Technology, where Hirose is a professor and Ono, a researcher at the Japan Marine Science and Technology Center, collaborates with another scientist. Both Hirose and Ono do similar work: squeezing materials, heated by a laser, between two diamonds in

a device called a diamond anvil. Both observe how the materials' properties change under extreme pressure, in conditions that mimic Earth's interior.

Both Hirose and Ono made important contributions to the discovery of the post-perovskite phase of the mineral MgSiO_3 . Yet when Hirose organized a conference on the topic in Tokyo last October, Ono wasn't invited to speak.

Hirose is outgoing, talkative and animated; he commanded the workshop with confident English that he had learnt during a



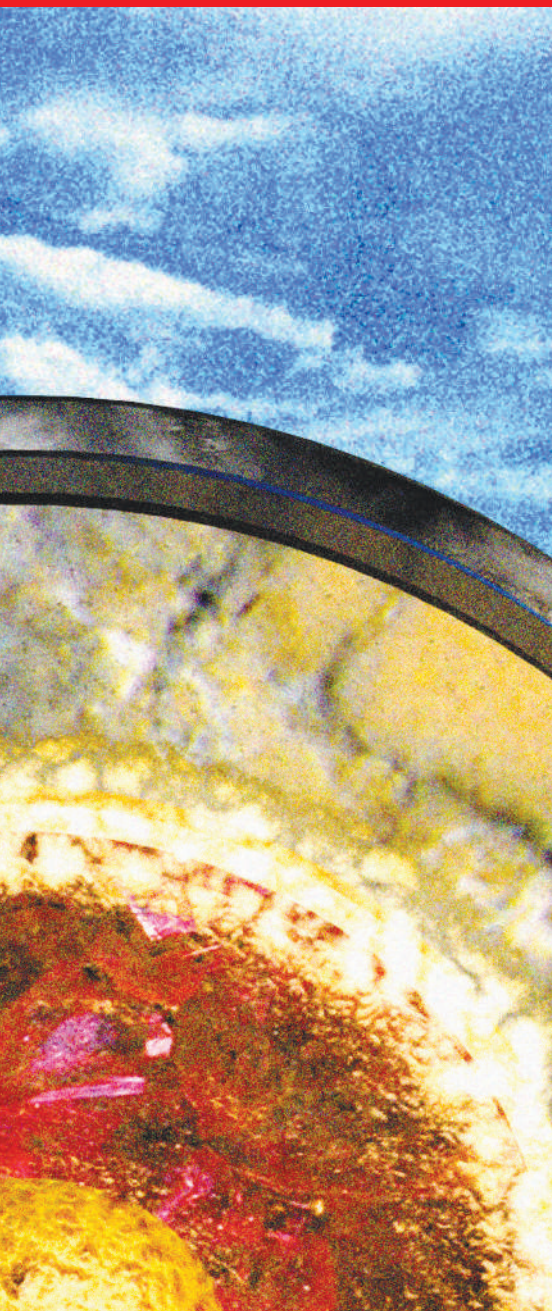
Kei Hirose: says Ono was "always working in secrecy".

postdoctoral fellowship at the Carnegie Institution in Washington

DC. Ono, who wanted to be interviewed in Japanese, is shy, serious and reserved.

But their disaffection is not just a clash of personalities. Each researcher has his own version of the discovery of post-perovskite.

Both men agree that advances in technology led the way. Based on a 1987 report, most researchers believed the perovskite form of MgSiO_3 was stable down to pressures near the bottom of the mantle, says Hirose¹¹. So there seemed little point in looking for post-perovskite. "Not many tried the experiment. It was very



C. DARKIN

planets. "As theorists we had been waiting for experimental clues," says Artem Oganov, a crystallographer at Swiss Federal Institute of Technology in Zurich. "This stimulated us."

After the first publication by Hirose's group, mineral physicists confirmed, both experimentally and theoretically²⁻⁴, the existence and structure of post-perovskite MgSiO_3 . And once researchers started looking at other minerals, such as germanates, which have germanium in place of silicon, they found it there as well. "It's surprising no one found it earlier," remarks Thomas Duffy, a geophysicist at Princeton University in New Jersey.

Post-perovskite seemed to explain many longstanding mysteries. For instance, simulations by Brodholt and his colleagues tackled the question of why a kind of seismic wave, called shear waves, suddenly accelerates at certain places within the D'' region. Shear waves also behave differently there than do compressional waves, another type of seismic wave. If the region contains post-perovskite, rather than perovskite, post-perovskite's properties would explain both observations⁵.

Similarly, post-perovskite can also explain why shear waves travel at different speeds within the D'' region depending on their direction, a phenomenon known as seismic anisotropy. The atomic orientation of post-

perovskite permits these waves to travel more quickly along certain crystallographic axes than others — giving rise to a pattern of anisotropy that can account for seismic observations in the D'' region^{6,7}.

Deep questions

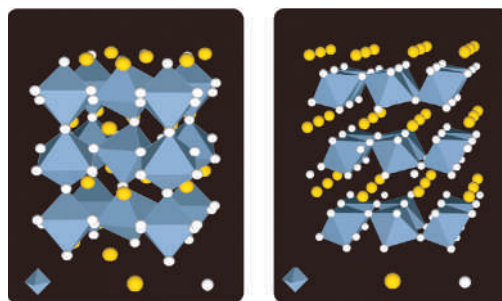
Such atomic-level studies illuminate bigger questions in Earth science — such as how heat and rock flow from deep in the Earth towards the surface, says Hirose. Figuring out how the lattice structure is deformed as perovskite becomes post-perovskite can tell scientists the

rate and direction of flow within the lower mantle.

Post-perovskite could help power the great engine of mantle convection in other ways. In some relatively cool regions of the mantle, the phase transition occurs at lower pressure, meaning at a shall-

lower depth. Compared with adjacent warm regions, the newborn post-perovskite is not only cooler but also more dense — giving it another reason to sink. "The phase transition is an additional pump helping to drive convection," says Thorne Lay, a seismologist at the University of California, Santa Cruz.

If post-perovskite is really so significant in inner-Earth dynamics, it would be a major coup for the newly discovered structure. But for now, post-perovskite often seems to provide more questions than answers.



Structures of perovskite (left) and post-perovskite.

K. HIROSE

difficult," says Hirose.

But by 2002, Hirose could carry out an experiment that he had envisaged more than five years earlier: subjecting a sample of perovskite MgSiO_3 to pressures and temperatures near those of the core-mantle boundary. He analysed the sample at the SPring-8 synchrotron facility in Hyogo, Japan, and found an odd diffraction pattern in the data.

For months, he tried to find a chemist to help out. "Chemistry professors told me there must be something wrong with my data — or something wrong with me," says Hirose. Finally Motohiko Murakami, a chemist colleague at Tokyo



Shigeaki Ono: thinks that Hirose treated him unfairly.

Institute of Technology, found time over his New Year's holiday to take

a look. Murakami identified post-perovskite in the sample.

But Ono says that a key step towards that discovery came more than six months earlier, when he found that the mineral haematite had a high-pressure phase transition — which yielded the same structure that would be found in MgSiO_3 by Hirose — and labelled it post-perovskite¹². "My objective was always MgSiO_3 ," he says.

Ono says Hirose was able to follow his work with haematite because researchers using SPring-8 are required to post data on a communal computer and an official beam-line notebook. "It wasn't really wrong, but it was

unfair, so I stopped collaborating with him," says Ono. Hirose says Ono "was always working in secrecy". The researchers' colleagues refuse to comment on the row.

The fallout may have been the inevitable consequence of a number of people converging on a discovery waiting to be made. By 1999, theorists had already formulated a phase transition with properties that would explain oddities in the D'' region of the inner earth. "It was remarkably accurate," says John Brodholt of University College London. "In other words, all it needed was experimental evidence."

D.C.

Under pressure:
diamond anvils
recreate the crushing
conditions found
deep within Earth.

K. HIROSE

For example, the studies of post-perovskite's role in seismic anisotropy all give different predictions of how the deformed structure would look. Researchers know they must fit the correct deformation pattern into their models — but they don't know which one is right.

Also missing is a precise understanding of the relationship between the pressure and temperature at which the phase transition occurs. The change is known to happen over a specific range of pressure and temperature, with higher temperatures requiring higher pressures. But pressure readings can vary by as much as 15 gigapascals depending on the laboratory equipment used in the experiments — translating to a difference in depth of 350 kilometres. "The pressure scale is the largest source of uncertainty," said Hirose last October, at a workshop on post-perovskite in Tokyo.

Spin doctors

Given that temperature varies greatly over the D" region, it is hard to know where post-perovskite might show up, argues Paul Tackley, a modeller at the Swiss Federal Institute of Technology in Zurich. His simulations⁸ suggest that a post-perovskite phase would be absent in relatively hot regions, which could be some 3,200 °C, or 1,000 °C hotter than other regions of the mantle at a similar depth. These are the regions thought to produce mantle plumes, the massive conduits of heat and rock that many think rise from deep in Earth to emerge as 'hotspot' volcanoes on the surface.

Large uncertainties also remain over one of the key questions in Earth dynamics — the extent to which Earth's core of liquid iron interacts with its mantle. Iron atoms can replace some of the magnesium in MgSiO₃; Wendy Mao, of the University of Chicago, Illi-

nois, has measured just how much iron post-perovskite can take up.

The answer, she found, is much more than any previously known lower-mantle material⁹. "Iron may have a dominant and very complicated role in post-perovskite," she says. Iron absorption may help account for the dramatic slowdown of seismic waves in certain pockets at the base of the mantle. But iron would also broaden the pressure range over which perovskite could turn into post-perovskite, making it harder to pinpoint the boundary where the phase transition occurs.

Such problems may arise from trying to understand a complex region with simple models of how post-perovskite behaves, says Lay. "The big issue is that the characterization of post-perovskite is all based on numerical models," he says. And for regions where other elements are intermixed in complex ways, "it is not clear that the numerical models are really relevant to Earth."

The details of the D" zone promise to get knottier before they get simpler. "After the initial euphoria that post-perovskite may explain everything, people are now taking a hard look at its properties," says Mao.

Despite the uncertainties, researchers are pushing into new areas. Shigeaki Ono, of the Japan Marine Science and Technology Center in Yokosuka is investigating MgSiO₃ as a possible explanation for one of Earth's mysteries. "Earth's rotation speeds up and slows down a few milliseconds over decade-long periods," says Ono. An electromagnetic coupling between the core and mantle might cause this, but if so the lower mantle must be highly electrically conductive. Perovskite does not have

high conductivity, so if the lower mantle is made mostly of perovskite — as many scientists believed — that theory cannot be right.

But Ono has found that a post-perovskite form of Al₂O₃ — an aluminium oxide that is better characterized than post-perovskite MgSiO₃ — is highly conductive. He argues that post-perovskite MgSiO₃ probably has many of the same properties — including a conductivity high enough to explain Earth's rotation discrepancy. "It's a new contribution of the post-perovskite," he says.

Other worlds

Some researchers have even looked at how post-perovskite MgSiO₃ might affect other planets. Renata Wentzcovitch, of the University of Minnesota, Minneapolis, and her colleagues have simulated how post-perovskite would behave inside gas giants or in rocky planets in other solar systems. In such high-pressure, high-temperature fields, they find, the MgSiO₃ would start to behave almost like a metal¹⁰.

Many scientists are excited by the possibility that on Earth there may be perovskite in some regions beneath the post-perovskite zone. This would explain seismic data that suggest shear waves change speed twice, first accelerating and then slowing. It could also constrain the temperature variations to a range of depths. "It would provide a kind of thermometer of the lower mantle and core," says Thomas.

For his part, Hirose is glad to have provided this useful, if confusing, discovery. Post-perovskite may not account for a large portion of the D" layer, but its existence is what matters.

"The proportions might not be huge," he says, "but with its location at the thermal and chemical boundary between the core and the mantle, it is going to help us understand crucial things about the Earth."

David Cyranoski is Nature's Asia-Pacific correspondent.

"After the initial euphoria, people are now taking a hard look at post-perovskite's properties."
— Wendy Mao

1. Murakami, M. *et al. Science* **304**, 855–858 (2004).
2. Oganov, A. R. & Ono, S. *Nature* **430**, 445–448 (2004).
3. Tsuchiya, T. *et al. Earth Planet. Sci. Lett.* **224**, 241–248 (2004).
4. Shim, S.-H. *et al. Geophys. Res. Lett.* **31**, L10603 (2004).
5. Wookey, J. *et al. Nature* **438**, 1004–1007 (2005).
6. Oganov, A. R. *et al. Nature* **438**, 1142–1144 (2005).
7. Merkel, S. *et al. Science* **311**, 644–646 (2006).
8. Hernlund, J. W. *et al. Nature* **434**, 882–886 (2005).
9. Mao, W. L. *et al. Proc. Natl Acad. Sci. USA* **101**, 15867–15869 (2004).
10. Umemoto, K. *et al. Science* **311**, 983–986 (2006).
11. Knittle, E. *et al. Science* **235**, 668 (1987).
12. Ono, S. *et al. J. Phys. Chem. Solids* **65**, 1527–1530 (2004).

BUSINESS

Crunch time for New Jersey lab

A global merger will put the long-suffering Bell Labs through the wringer yet again. But could joint ownership benefit the world-famous facility in the long run? Kurt Kleiner investigates.

Bell Labs has suffered a few setbacks in recent years, as its owner, telecoms company Lucent, has cut back on basic research. Now, the 1,100 scientists and engineers at the lab in Murray Hill, New Jersey, face another upheaval: they've been bought out by a French telecoms company.

Alcatel's €10-billion (US\$12-billion) take-over of Lucent will create the world's biggest maker of telecommunications equipment. It could also create even more pressure to cut back basic research at Bell Labs. But despite company statements that 10% of the combined 88,000 workforce are likely to lose their jobs, some observers say the merger could strengthen Bell Labs in the long run.

"The concern I always had with Lucent and Bell Labs was that the company couldn't afford the research that Bell Labs was doing," says John Slack, an analyst with investment-research company Morningstar in Chicago. "The merged companies can throw in a lot more research dollars."

Both Paris-based Alcatel and Lucent, based in New Jersey, supply wireless and land-based telecommunications equipment and services. If, as expected, they obtain approval from their shareholders and governments, they will join forces in six to twelve months, to form a Paris-based company that has yet to be named.

The two companies say that they expect to save €1.4 billion annually, mainly by laying off staff. Although billed as a merger of equals, Alcatel, with sales of €13.1 billion and 58,000 employees, is almost twice as big as Lucent, with 30,500 employees and €7.7 billion in sales (see graphic). And the French company's shareholders will own 60% of the new company.

Scientists at Bell Labs invented the transistor, discovered the cosmic microwave background radiation left over from the Big Bang, and won six Nobel Prizes, but have, more recently, weathered years of cutbacks. Since it was spun off from its former parent, AT&T, to become part of Lucent in 1996, it has lost two-thirds of its staff, and focused more on product development than on basic scientific research



Best of pals: Serge Tchuruk, chair of Alcatel, left, and Lucent's chief executive Patricia Russo expect their merger to yield the world's top telecoms supplier.

(see *Nature* 440, 146–147; 2006).

"It has lost a lot of its lustre and size from its heyday," says Jay Pultz, an analyst at Gartner Research in Stamford, Connecticut, and a former Bell Labs systems engineer. "It still does very good research, but the number of people involved in that, and the number of research areas that that covers, is much more limited." He estimates that only about 5% of Bell Lab's total research and development budget now goes on basic research, and that about 200 researchers are involved in it.

The merger could put even more pressure on that basic-research component, says Narain Gehani, a computer scientist at the New Jersey Institute of Technology, former Bell Labs researcher and author of *Bell Labs: Life in the Crown Jewel*. "A company that makes telecom-

munications equipment without leading-edge research doesn't have much of an edge. So Bell Labs is critical. But it cannot do unfettered research," says Gehani. "In my view what will happen is that it will be oriented more and more towards the business."

The merged company might also hesitate to cut Bell Labs for political reasons, says Niel Ransom, former chief technology officer of Alcatel and a former Bell Labs researcher. "Bell Labs brings a lot of political value. They want to be able to show that as they become a more international company, it will not have a negative effect on research," he says.

Bill Price, a Lucent spokesman, said that although no specifics have been announced, layoffs will be spread across all parts of the company. On a *pro rata* basis, that could mean up to 2,600 scientists and engineers losing

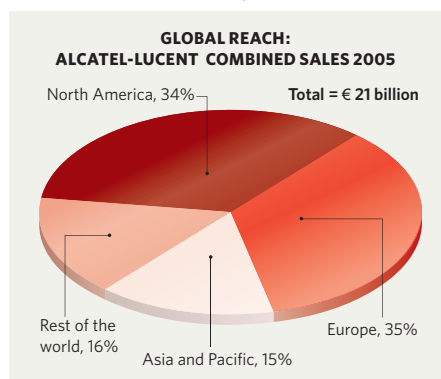
"The merged companies can throw in a lot more research dollars."

their jobs worldwide. Ransom thinks that product-development teams doing duplicate work will absorb the steepest job cuts. For instance, engineers working on computer networking devices could lose their jobs, as both companies make this equipment. But

Lucent engineers working on advanced wireless communication technologies such as CDMA (Code Division Multiple Access) don't have any counterparts at Alcatel, and should, Pultz says, be able to breathe more easily.

Government approval for the merger could be held up because the companies do classified work for their respective governments. But Alcatel has announced plans to spin off some of its defence-related work. And Lucent says it would set up a US-controlled subsidiary to handle classified work.

It will take a year or two to see how things shape up. But Pultz hopes the new company sees the importance of supporting research and development. "At some point, if you don't make these long-term investments, it will come back to haunt you," he says. "Even in this era of having to get more profitable and prove themselves to Wall Street, you can't be that short-sighted." ■



Everyone listed on Dolly paper met established criteria for authorship

SIR — Your Special Report “Credit where credit’s due” (*Nature* 440, 591–592; 2006) is critical of the extent of Ian Wilmut’s claims to credit for Dolly the cloned sheep, as well as of authorship issues connected with the publication of her creation (*Nature* 385, 810–813; 1997).

Wilmut has modestly stated that his role in the research was mainly a supervisory one, and that others, principally Keith Campbell, deserved more of the intellectual credit. These comments, together with those made by a disaffected previous member of Wilmut’s group, were seized upon irresponsibly by *Nature*.

As someone intimately involved at all stages of the Dolly project, I am confident that authorship was properly assigned and that due credit has been given.

Dolly came into existence as a result of Wilmut’s efforts to attract funding for nuclear-transfer research and his recruitment of excellent staff to pioneer innovative techniques. These amply justify his public standing as the person most clearly associated with Dolly, an outcome that has unfortunately been confused with apportionment of credit.

Dolly resulted from the combined efforts of scientists, veterinary surgeons, farm managers and farmhands, most of whom were not listed as authors.

To qualify as an author, one should have made key intellectual and/or novel technical contributions — inclusion of all, irrespective of the nature and extent of their contribution, can devalue the currency of those who are chiefly responsible for a project’s success.

All the individuals listed on the Dolly paper, including Wilmut, met established criteria for authorship. The technical breakthrough that made Dolly possible was published by Wilmut and his team in 1996 (*Nature* 380, 64–66; 1996), one year before the Dolly paper. But the content of scientific papers that chronicle major discoveries is not always aligned with the impact those papers make and, unsurprisingly, of the two, the Dolly publication took the lion’s share of public attention.

Authorship is an important issue for the scientific community and for scientific journals. Important steps are being taken to ensure that authorship is properly credited, as you mentioned in your Special Report.

But authorship issues are not limited to papers about cloning and stem cells. Such a focus could, in the present climate, provide further ammunition to those wishing to

smear the integrity of all involved in this important area of scientific research.

Alan Colman

ES Cell International Pte Ltd,
11 Biopolis Way, #05-06 Helios,
Singapore 138667

We reject the suggestion that we were irresponsible in factually reporting the dispute relating to Dr Wilmut in the context of an account of disputes about credit; in no way, explicitly or implicitly, did we suggest that he was at fault. Editor, *Nature*.

It’s incredible how often we’re surprised by findings

SIR — How often do natural scientists admit surprise at their findings? More frequently than social scientists or non-scientists? I searched for words indicating surprise among 30,133,141 abstracts of English-language scientific papers indexed in the Science Citation Index (SCI) and 8,151,087 English-language academic articles in the Social Sciences and Arts & Humanities citation indices. I compared the frequencies of words appearing in the SCI with their frequencies in the other indices and, separately, with their occurrence in general writing, as recorded in the Brown Corpus of Standard American English (edict.com.hk/lexiconindex), by testing the significance of log-odds ratios (all were found to be significant at $P < 0.001$).

The study of nature does indeed seem to surprise us. The odds of finding in abstracts of scientific research papers a result or conclusion described as ‘surprising’, ‘unexpected’, or ‘unusual’ are an order of magnitude greater than in standard language and several times greater than in non-science academic abstracts.

The word ‘surprising’ appears 12 times more frequently in the natural sciences than in standard English and 1.3 times more frequently than in social sciences, arts and humanities. The word ‘unexpected’ appears 39 times and 2.2 times more frequently in the natural sciences than, respectively, in standard English and in non-science academic writing.

In contrast, words such as ‘happy’, ‘unhappy’ or ‘ugly’ occur with frequencies that are expectedly lower in the natural sciences than elsewhere (further details of this research are available on request from jasienski@post.harvard.edu).

Although natural phenomena can indeed sometimes be surprising if they are against our expectations, being ‘surprising’ is not an inherent quality of nature. Does scientists’ use of this term in their publications truly represent genuine surprise at their results?

One might think that academic machismo or realism would cause scientists to downplay their surprise, but, on the other hand, overstating the level of astonishment may occur when striving for media attention.

Michał Jasienski

Nowy Sacz Business School –
National-Louis University, Zielona 27,
33-300 Nowy Sacz, Poland

Funding should recognize outcome, not income

SIR — I was disappointed to read in your Editorial “Brown’s budget briefing” (*Nature* 440, 581; 2006), discussing a future replacement for the UK Research Assessment Exercise (RAE), that you consider “external research income” as “a reasonable basis for departmental funding”.

I hope it is still true to say that no major scientific prizes or seats at the high tables of science have been awarded on the basis of an individual’s research income. Although external research income has been a significant — some might say disproportionate — factor in previous RAEs, it should not become a substitute for scientific excellence.

For scientists, the only significant research outcome is the science and, primarily, its communication through publication. Scientists are all too aware of instances of poor accountability in the spending of highly competitive research grants — the result of grant-giving bodies supporting the idea and not the individual.

A scientific meritocracy based upon success in gaining research funding is wholly reversible. But there are no instances of scientists whose success has been measured by their output, on the other hand, being asked to leave their seats at science’s high table, their Nobel prizes in envelopes marked “Return to sender”.

In other words, the measure of scientific success is scientific output — the elucidation of new knowledge and its dissemination through publication — and not science income.

Awarding science funding primarily upon the basis of the latter will only ensure a decline in the former.

Christopher Exley

Birchall Centre for Inorganic Chemistry
and Materials Science,
Keele University,
Staffordshire ST5 5BG, UK

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. They should be signed by no more than three authors; preferably by one. Published contributions are edited.

BOOKS & ARTS

Where are the switches on this thing?

Asking the big questions in neuroscience.

23 Problems in Systems Neuroscience

edited by J. Leo van Hemmen & Terrence J. Sejnowski

Oxford University Press: 2006. 530 pp.
£49, \$79.95**Kevan A. C. Martin**

David Hilbert, in his opening address at the International Congress of Mathematicians in Paris in 1900, presented his colleagues with 23 problems whose investigation he thought would provide the major advances in mathematics in the twentieth century. Although about half of the problems remain unsolved, history shows that mathematicians rose splendidly to the challenge.

Neuroscience has a rather briefer history than mathematics, but Leo van Hemmen and Terry Sejnowski felt that it was nonetheless mature enough for them to organize a meeting on 'Problems in Neuroscience' a century after Hilbert's address. This printed version of their meeting, *23 Problems in Systems Neuroscience*, has taken six years to arrive, but it is not too late and certainly not too little. In the place of one Hilbert are 40 problem-posers who have collectively contributed the 23 chapters, grouped into sections that sum up 5 current concerns: How have brains evolved? How is cerebral cortex organized? How do neurons interact? What can brains compute? How are cognitive systems organized? With such an attractive list of topics, this book is sure to find a wide audience at every level of interest, from lay readers to students and academics.

For most of the contributors, the operations of the cerebral cortex pose the principal problems to be solved. It is refreshing, then, to feel the frank blast of Gilles Laurent's irritation on page 1 when he asks: "Why this obsession with cortex?" From his locust-brain perspective, it seems to Laurent that "most scientists act as if King Cortex appeared one bright morning out of nowhere, leaving in the mud a zoo of robotic critters, prisoners of their flawed designs and obviously incapable of perception, feeling, pain, sleep, or emotions, to name but a few of their deficiencies. How nineteenth century!" This opening salvo is exactly the energetic and uncompromising tone required to sweep the reader onwards through the sparkling debates that fuel this book and to help define the most important questions in neuroscience.

The writing frequently made me smile often it comes close to realizing the uncommon view of an old French mathematician retold by Hilbert: "A mathematical theorem should not be considered complete unless you have made it so clear that you can explain it to the first man (or woman) whom you meet on the street."

Like Hilbert, the authors assume that their problems can and will be solved. However, unlike mathematical problems, those of systems neuroscience will not be solved by pure reason, but will require the integration of experiment and theory, and perhaps even a new science that incorporates the subjective. In this context it would be interesting to know how many of the authors originally trained as neuroscientists or even biologists. I suspect, as I suspect, card-carrying neurobiologists are in the minority, then this book exposes the deep, dark secret that neuroscience has become one of the most pathetic of modern scientific disciplines as it should be, because neuroscience faces the most intractable questions in modern science and we need all hands on deck if we are to find a safe passage through the problems raised here.

However, most contemporary experimental neuroscience is done at the molecular and cellular level, where most of the hard problems discussed in the book need not be faced. Although the book does not consider molecular and cellular techniques, it seems increasingly clear that they will be central to the approaches required to solve many of the systems-level problems.

Similarly, although Cathy Carr and her colleagues do discuss some structurally significant adaptations of neurons in the auditory system, questions of the relevance of structure to systems neuroscience are hardly addressed here at all. Jim Watson and Francis Crick emphatically demonstrated that some otherwise intractable problems in biology can be cracked by solving structures. But structural neuroscience has limited sex-appeal for the editors of leading journals. This, together with the granting agencies' blind-spot for structure, has resulted in a lack of knowledge about the fine structure of the nervous system



M. KULKA/ZEFA/CORBIS

Neuroscience would be much easier if we had a detailed circuit diagram of the brain.

of every model animal, *Caenorhabditis elegans* excepted, and this remains a major impediment to progress. Sejnowski points to a related problem when he asks: "What are the projective fields of cortical neurons?" The receptive fields of neurons, which result from the convergent input of many neurons, have been well studied for decades. The projective fields — the influence of the output of a neuron on its many targets — are largely unknown but would be much better understood if we had a detailed circuit diagram of the cortex and its related structures.

Several of the authors have produced entertaining and provocative critiques of the state of current techniques and thinking. A sharp intake of breath will no doubt be heard from many readers of the piece by Bruno Olshausen and David Field. They make the seemingly outrageous claim that we are ignorant of 85% of the functions of the primary visual cortex, area V1, the cortical area that most textbooks present as being essentially solved. But this provocative view does not go unsupported by experimentalists for, as Amiram Grinvald and

colleagues show, their techniques for recording large numbers of neurons in the primary visual cortex with fine spatial and temporal resolution inevitably drive new concepts of cortical function. Fortunately, discussions of the visual cortex do not dominate the book. Perhaps this is because the key themes that run through it are questions of what the neural code is, and how time is represented in the brain, and both questions apply equally well to the brains of creatures great and small. These two questions are also considered in virtuoso solos by Carl van Vreeswijk and Andreas Herz.

A prize for the wittiest title goes to Larry Abbott who asks: "Where are the switches on this thing?" He points out that the problem of self-regulated switching is solved for digital computers, which are the most complex machines we have built, but when it comes to the flow of information through a brain we know only that there are ways of biasing the flow by mechanisms such as neuromodulation or inhibition. Thus, with a motivation certain

to warm the heart of every invertebrate physiologist, he recommends that we study any reasonable candidate mechanisms for switching in a brain.

Hilbert stated that "the importance of a scientific work can be measured by the number of previous publications it makes it superfluous to read". Answers to some of the questions raised here will no doubt make the literature reviews of most PhD theses shorter, but what will we really understand even if all 23 problems in systems neuroscience are solved? This is the most difficult question of all, and its answer has a significance that goes far beyond even the ambitious goal of understanding the workings of a brain. Hilbert's vision for mathematics was carved on his tombstone — "Wir müssen wissen, wir werden wissen" ("We must know, we shall know") — but for those who study brains, even if we must, will we? ■

Kevan A. C. Martin is at the Institute of Neuroinformatics, ETH/University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland.

passed, we realized that we are less and less special. Today, we see ourselves as insignificant in the context of the whole Universe. The copernican revolution relegated and redefined Earth as just another planet, and made the Sun the hub of the Universe. Then astronomers showed that the Sun is not even at the centre of the Milky Way, and eventually it became clear that there are billions of other galaxies, which made Earth seem trivial.

The problem with becoming increasingly insignificant was appreciated as far back as the seventeenth century by the French philosopher Blaise Pascal: "I feel engulfed in the infinite immensity of spaces whereof I know nothing and which know nothing of me. I am terrified... The eternal silence of these infinite spaces alarms me."

The existence of dark matter only further relegated humanity, as Primack pointed out in 1984: "Yet another blow to anthropocentricity: not only is man not the center of the universe physically (as Copernicus showed) or biologically (as Darwin showed), it now appears that we and all that we see are not even made of the predominant variety of matter in the universe!"

But Abrams and Primack argue that humans still hold a central and special position in the Universe, perhaps not geographically but in many other ways. For example, we are special because we are made of the rarest material in the Universe, namely large atoms. Also, we live at a central time, because most nearby galaxies are past their violent youths but are not yet senescent. And we live at the midpoint of our planet's life, which is a few billion years old, and which has a few billion more years before it will be roasted by our Sun swelling into a red giant. And humans have a reasonably central size, roughly halfway between the smallest length scales (10^{-33} cm) and the distance to the cosmic horizon (10^{28} cm).

Seeking meaning in the void

The View from the Center of the Universe: Discovering Our Extraordinary Place in the Cosmos

by Joel R. Primack & Nancy Ellen Abrams
Riverhead Books: 2006. 400 pp. \$25.95

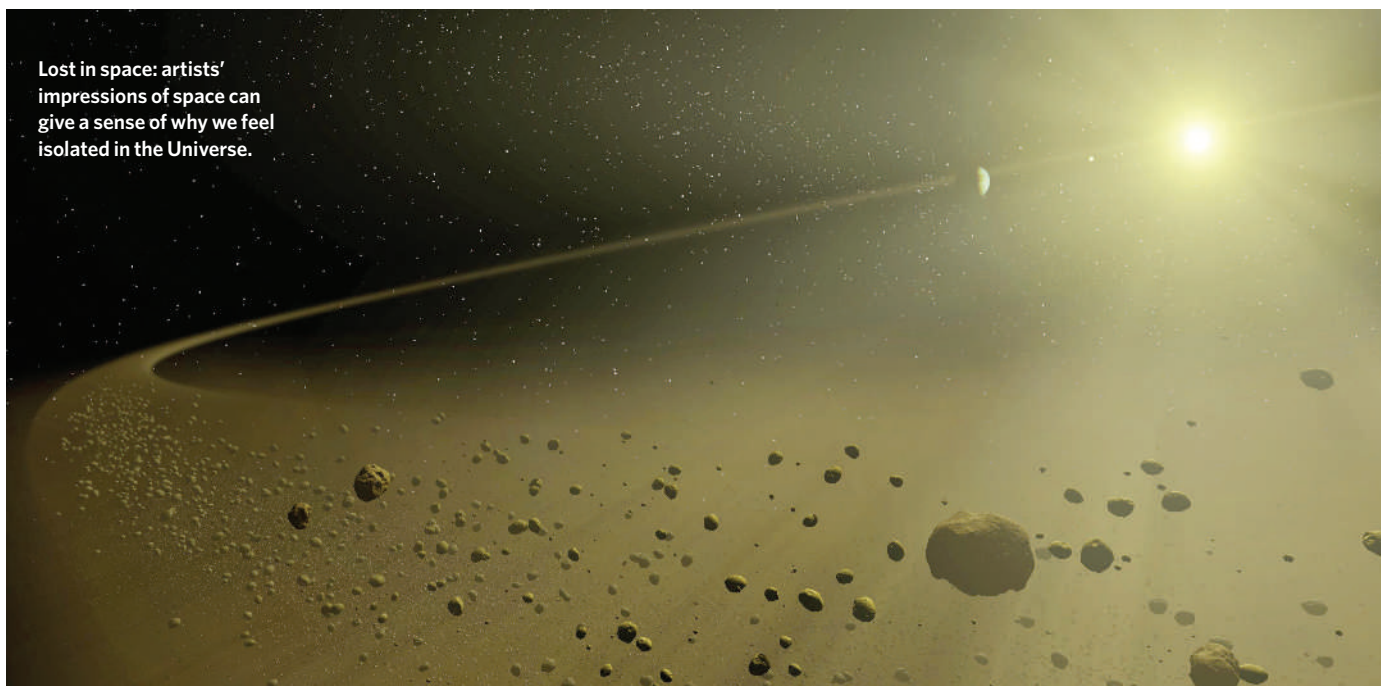
Simon Singh

The married couple of philosopher Nancy Ellen Abrams and cosmologist Joel R. Primack are uniquely placed to discuss how our understanding of the Universe affects how we perceive our role in it. The ancient creation myths

provide comfort and meaning, but they are fantasies. In contrast, modern cosmology offers a glimpse of reality but leaves many people cold. In *View from the Center of the Universe*, Abrams and Primack challenge themselves to try and get the best of both world views.

In the distant past, we convinced ourselves that we had a special place in the Universe. Geographically we were at the centre of space, with everything revolving around us, and biologically we thought that humans were an exceptional creation. But as each century

Lost in space: artists' impressions of space can give a sense of why we feel isolated in the Universe.



NASA/JPL-CALTECH

The book builds a good case as to why science should make us feel special, but sometimes the language, which wallows in mysticism, is irritating. We encounter the Pyramid of All Visible Matter, topped by the all-seeing eye, and parallels are drawn between cosmology and the Kabbalistic Order of Creation. There are also strange links made between physics and politics, such as the section that uses the laws of gravity and circular motion to explore the question of wealth distribution.

Abrams and Primack work hard to craft a

view of science that might allow us to connect with the Universe, but it is risky to mix science with New Age jargon, particularly when there is a risk of confusing non-scientists. Films such as *What the Bleep Do We Know!?* and dozens of pseudoscience books twist the bizarre laws of quantum physics to support all sorts of unscientific nonsense, and readers intrigued by such wacky notions will only have their ideas consolidated if they read about the Sovereign Eye, Abrams and Primack's mystical label for the conditions that give rise to intelligent life.

Although I have doubts about some of the language that the authors use to try and reconcile science and mysticism, I respect their efforts and some of their ideas. It is admirable that they have considered this problem worthy of discussion, even if they do not yet have all the answers. As Abrams and Primack point out, Einstein supposedly said: "Problems cannot be solved at the same level of awareness that created them."

Simon Singh is a science writer and the author of *Fermat's Last Theorem* and *Big Bang*.

Sympathy for the devil

Tasmanian Devil: A Unique and Threatened Animal

by David Owen & David Pemberton

Natural History Museum/Allen & Unwin:
2006. 240 pp. £12.99/\$24.95

Matthew J. Phillips

Two representations have dominated public perceptions of the largest living marsupial carnivore, the Tasmanian devil. One is the voracious, hurricane-like innocent savage Taz of *Looney Tunes* cartoon fame. The other, familiar in nineteenth- and twentieth-century rural Tasmania, is the ferocious predator and scavenger that wantonly kills livestock — and perhaps even people, should they become immobilized in the wilderness at night. Devils can take prey nearly three times their size and eat more than a third of their body weight in a sitting. Even so, it is hard to imagine how this species, being only slightly larger than a fox terrier, could be so maligned in name and image.

In *Tasmanian Devil*, David Owen and David Pemberton delve into devil biology to convey the true nature of the beast once known to science as *Sarcophilus satanicus* (now *S. harrisii*). Fact and fiction are teased apart with sound science and tempered speculation. The devil's behaviour and physical appearance are explained in terms of its unique ecological position as a solitary nocturnal predator that relies heavily on communal scavenging. Its larger cousin, the thylacine (*Thylacinus cynocephalus*), is now extinct, so the devil's present ecological interactions and selection pressures may differ somewhat from those under which it evolved. This makes the authors' comparisons with placental analogues — the ratel (honey badger), wolverine and hyena — particularly instructive. Although a useful starting point for those with an academic interest in the Tasmanian devil, this book, with its well chosen photographs and historical illustrations, has far wider appeal.

The humour and tragedy associated with early European settlers' misunderstanding of the devil are neatly woven together, and the authors' arguments that the devil is not a rural



P. A. SOUTERS/CORBIS

Just misunderstood? Science has shown that the Tasmanian devil isn't so satanic after all.

menace are appealing. But I wonder whether the use of anecdotal evidence to lay the blame for poultry and trap raiding on the even rarer spotted-tailed quoll (*Dasyurus maculatus*) only extends the tyranny of prejudice.

Peripheral connections to the devil story provide light relief. Particularly well fleshed out is the link between Theodore Flynn, who studied devil reproductive anatomy, his actor son Errol, who dubbed himself 'the Tasmanian devil', and Errol's employer, Warner Brothers, who have profited immensely from Taz cartoons and merchandising.

The inclusion of a wide array of reports and newspaper articles provides the reader with access to a mostly bygone mood of malevolence towards the devil, as well as to the voices that began to change this attitude. It is particularly sad that having survived being shot, poisoned and trapped for bounties, and finally winning considerable public affection, devils are now succumbing to devil facial tumour disease. The authors relate the few clear facts

about this hideous affliction, which seems to spread through biting and is devastating devil populations across much of Tasmania. It is unknown whether the disease is an old foe or whether its origins lie elsewhere, for example in the accumulation of anthropogenic carcinogens. At this and other points of uncertainty I was left wondering what the Tasmanian aboriginals could have told us about the devil, had misunderstanding, persecution and disease not led to their own demise.

The authors have succeeded in demystifying the Tasmanian devil and reveal a fascinating creature; we would be much poorer without it. Nevertheless, if you were to follow some raucous screams through the dark Tasmanian night and came upon half a dozen of these stout, black marsupials gorging on the carcass of a cow with their bone-crunching teeth, you might still think they were devils indeed.

Matthew J. Phillips is at the Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.

NEWS & VIEWS

LANGUAGE

Startling starlings

Gary F. Marcus

Recursion, once thought to be the unique province of human language, now seems to be within the ken of a common songbird — perhaps providing insight into the origins of language.

Man the tool-maker. Man the cultural animal. Man the mimic. It's tempting to summarize the differences between humans and other species in a concise phrase, but most posited differences have turned out to be overstated. Chimpanzees and gorillas use sticks to fish for termites; orangutans use sticks for autoerotism. And many of these capacities seem to be culturally mediated; they are transferred from one primate to the next by illustration and observation, rather than learned afresh by trial and error¹.

The report by Timothy Gentner and colleagues on page 1204 of this issue² challenges one more putatively uniquely human adaptation: the capacity to recognize complex 'recursive' structure. Gentner *et al.* showed that at least one non-human species, the European starling (Fig. 1), can be trained to acquire complex recursive grammars such as the A^nB^n language (in the case of the starling, *rattle rattle warble warble*; see below).

Recursion, or self-embedding, is without question a hallmark of human language. For example, one can take a phrase such as *love conquers all* and embed it in a frame such as *X knows Y*, yielding, say, *Chris knows love conquers all*. The output of that process can then be fed back into the *X knows Y* frame, yielding, say, *Terry knows Chris knows love conquers all*. Embedding also makes relative clauses possible, as in the bracketed part of *the paperback [on the coffee table] is hilarious*. Marc Hauser, Noam Chomsky and Tecumseh Fitch³ have speculated that recursion might be unique to humans — and perhaps even the only contribution to language that is human-specific. Consistent with this, Fitch and Hauser⁴ found that cotton-top tamarin monkeys could not distinguish the A^nB^n language from an ostensibly similar language $(AB)^n$ that need not be constructed recursively.

The A^nB^n language (see Fig. 1 of the paper² on page 1204) is generally assumed to be recursive because new sentences can be formed by successive insertion into the frame AXB , for example AB , $AABB$, $AAABBB$ and so on. Gentner and colleagues² rewarded European starlings for pressing a bar in response to



D. D. BALECKAITS

Figure 1 | No bird brain. The European starling, *Sturnus vulgaris*, which Gentner *et al.*² show is capable of recognizing complex grammar.

A^nB^n strings of starling-generated sounds, such as *rattle rattle warble warble*, and withheld the reward for responses to the $(AB)^n$ grammar (and vice versa for another group of starlings). Although learning was not instantaneous, nine of eleven birds eventually (after 10,000–50,000 trials) learned to discriminate reliably between the two grammars, succeeding where the monkeys had failed. An extensive series of control comparisons strongly suggests that the ultimately acquired grammar is robust. Notwithstanding some minor worries⁵, this is strong evidence that humans are not alone in their capacity to recognize recursion.

What explains the discrepancy between the starlings' success and the tamarins' apparent failure? One possibility is methodological: Fitch and Hauser⁴ tested whether tamarins could acquire A^nB^n spontaneously from a relatively brief exposure, whereas Gentner *et al.*² asked whether starlings could acquire similar

structures from considerably longer exposures, enhanced with positive feedback, in a more active task. Only further experimentation can clarify whether tamarins' apparent inability to recognize the A^nB^n structure is context-specific or genuinely absolute.

If the split between tamarins and starlings does prove robust, further comparative work will clearly be necessary. Can other varieties of birds that don't (in contrast to starlings) naturally acquire new songs also acquire self-embedded structures? Are humans alone among primates in their capacity to do so? Might the capacity for recursion be general across great apes, even if it were absent in monkeys? An intriguing possibility is that the capacity to recognize recursion might be found only in species that can acquire new patterns of vocalization, for example songbirds, humans and perhaps some cetaceans.

Whether or not tamarins can be prodded

into recognizing recursion, the analogous human capacity seems robust. Humans are quick to notice recursion and are able to do so without explicit reinforcement; perhaps most importantly, they can generalize recursive structures broadly. Starlings have thus far been shown to be able to extend $A^n B^n$ only to new sequences of familiar sounds. Humans can clearly go further; once you recognize the pattern in $AABB$ and $AAABBB$, it is a trivial matter to extend that pattern to new vocabulary (for example, $CCCDDD$ or $JJJKKK$). Taken together with the tamarins, there actually seems to be a three-way split: some species may generalize recursion only to items that have already been instantiated in a given pattern; some species can generalize recursion freely to newly acquired vocabularies (arguably the essence of human language); and some species apparently cannot recognize recursion at all.

The “abstract computational capacity of language”³ may consist not so much of a single innovation as a novel evolutionary reconfiguration of many (perhaps subtly⁶ or even qualitatively⁷ modified) ancestral cognitive components, genetically rejigged into a new whole. Contemporary research suggests that the human brain contains few if any unique neuronal types, and few if any genes lack a significant ancestral precedent⁸. At the same time, humans show much continuity with their non-speaking cousins in dozens of ways that might contribute to language, including mechanisms for representing time and space, for analysing sequences, for auditory analysis, for inhibiting inappropriate action, and for memory.

None of this challenges Chomsky’s long-held conjecture⁹ that children are innately endowed with a universal grammar — a set of mental machinery that would lead all human languages to have a similar abstract character. But that shared abstract character may have as much to do with our lineage as vertebrates as with our uniquely human innovations. In Charles Darwin’s immortal words, “throughout nature almost every part of each living being has probably served, in a slightly modified condition” in some ancestor or another. ■

Gary F. Marcus is in the Department of Psychology, New York University, 6 Washington Place, New York, New York 10012, USA.
e-mail: gary.marcus@nyu.edu

SUPERCONDUCTIVITY

Quantum stripe search

Jan Zaanen

Do quantum stripes exist or not? Further indirect evidence for this controversial behaviour of electrons in high-temperature superconductors comes from measurements of atomic-lattice vibrations.

Since 1995, the ‘stripe wars’ have been raging in the demesne of high-temperature superconductivity. This fierce conflict, fought with the highest-calibre weapons of experimental physics, has its antecedent in a hotly contested claim about the way electrons behave in the copper oxide materials notoriously used as high-temperature superconductors. Supposedly, they form highly organized patterns called quantum stripes — but only on the picosecond timescale, so the patterns average away over longer periods through the electrons’ constant quantum dance.

The dispute has lasted so long only because it has proved very hard to nail down such genuine quantum behaviour. In this issue, however, Reznik *et al.* (page 1170)¹ present further evidence in support of quantum stripes. They show that the collective vibrations of the atomic lattice of certain superconducting copper oxides behave in a manner that is hard to explain — unless one assumes that motions characteristic of the presence of quantum stripes are shaking the ion lattice. So is the end of hostilities in sight?

It is an everyday experience that, in a many-body system, collective behaviours emerge that are utterly unrelated to the behaviours of the objects that make it up. The evolution of a nation’s economy over time, for example, is difficult to predict from the often conflicting motivations of the constituent human members. The same principles apply in quantum physics. But the exact rules that govern quantum emergence are poorly understood; uncovering them is a core business of modern physics.

Conventional superconductors — those operating only at temperatures very close to absolute zero — demonstrate only a minimal form of quantum emergence. In such materials, interactions between electrons diminish at low temperature, and the macroscopic electron system turns into a near-ideal (non-interacting) quantum gas of ‘quasi’-electrons. A small residual attractive interaction binds these quasi-electrons in pairs, which in turn collapse to a single quantum state in the process known as Bose–Einstein condensation.

Although this theory of superconductivity, known as the Bardeen–Cooper–Schrieffer (BCS) model, gets everything right for conventional superconductors, it explains hardly anything in high-temperature superconductors. The discovery 20 years ago of this unusually sturdy form of superconductivity raised the curtain on a drama of wider relevance: the

huge numbers of strongly interacting electrons in the copper oxide layers were plainly showing an unknown kind of collective quantum physics (for an overview, see ref. 2). In 1995, it was discovered³ that small changes in the crystal structure of high-temperature superconductors can cause superconductivity to disappear, with a peculiar ‘static stripe phase’ taking over (Fig. 1). Here, strong interactions and quantum motions work together to form patterns of electrons moving in serried ranks. Domains in which the electrons come to a complete standstill separate these ‘rivers of charge’ (Fig. 1b).

The existence of static stripes, initially contentious, is now generally accepted. But quantum stripes are more radical and controversial. Members of the quantum-stripe faction hold that stripes are, in fact, always present. When a material superconducts, the stripes do not disappear; rather, a quantum-mechanical superposition of countless disordered stripe states forms (Fig. 1a), in such a way that the overall state corresponds to that of a superconducting quantum liquid (for a mathematical proof of principle, see ref. 4).

So how can we nail down quantum stripes experimentally? Consider Erwin Schrödinger’s oft-cited cat hidden in a sealed box (Fig. 1c). Classically, the cat must be either dead or alive, but quantum mechanically it is in a superposition of dead and alive states. In the quantum state, the cat fluctuates back and forth between alive and dead states. This fluctuation takes a finite time. So, by taking snapshots quickly enough, one can see either a dead or a live cat. The same scheme works equally well with superpositions of countless many-electron configurations. Here, every quantum configuration takes the role of a classical ‘dead or alive’ state (Fig. 1c).

The fluctuation time of the quantum stripes is in the picosecond (10^{-12} second) range, and the problem for the experimentalist is how to grab a picture of complicated spatial electron patterns in so little time. One way to do this is to observe the change in kinetic energy of neutrons that scatter off the material inelastically. Since 1995, such neutron-scattering experiments have added to the body of evidence supporting the case for quantum stripes⁵. The drawback is that these studies relied on information about the direction of the electron spins that was open to alternative interpretations, including some compatible with the conventional BCS picture (see ref. 6 and references

- Whiten, A. *Nature* **437**, 52–55 (2005).
- Genter, T. Q., Fenn, K. M., Margoliash, D. & Nusbaum, H. C. *Nature* **440**, 1204–1207 (2006).
- Hauser, M. D., Chomsky, N. & Fitch, W. T. *Science* **298**, 1569–1579 (2002).
- Fitch, W. T. & Hauser, M. D. *Science* **303**, 377–380 (2004).
- Kochanski, G. *Science* **303**, 377–380 (2004).
- Pinker, S. & Jackendoff, R. *Cognition* **95**, 201–236 (2005).
- Marcus, G. F. *The Birth of the Mind: How a Tiny Number of Genes Creates the Complexities of Human Thought* (Basic Books, New York, 2004).
- Fisher, S. E. & Marcus, G. F. *Nature Rev. Genet.* **7**, 9–20 (2006).
- Chomsky, N. *Language and Mind* (Harcourt, Brace & World, New York, 1968).

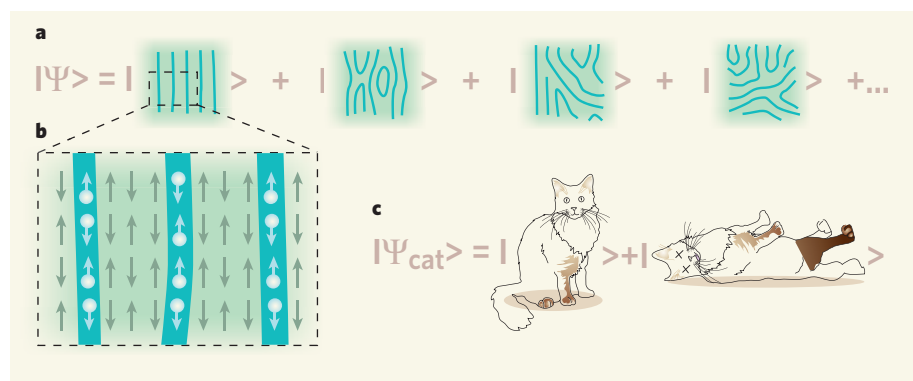


Figure 1 | Hunting the stripes. Electrons in the copper oxide planes of high-temperature superconductors sometimes form non-superconducting static stripe patterns. **a**, These stripes are believed to survive in the superconducting state as part of a quantum superposition of countless disordered stripe states that forms an overall featureless superconducting quantum liquid. **b**, Close up, the stripes are seen to consist of delocalized electrons ('rivers of charge') separated by domains of localized electrons showing a characteristic spin order. **c**, The principle of quantum superposition by which the stripe states are hidden from view in the superconducting phase is also the source of uncertainty over the fate of Schrödinger's infamous cat. Capturing the quantum cat or the quantum stripes in a definitive state requires snapshots taken on a timescale that is short compared with the timescale of quantum fluctuations between the superimposed states (dead or alive; stripy or non-stripy). By exploiting the coupling between high-frequency lattice and collective-stripe vibrations, Reznik *et al.*¹ provide a glimpse of quantum stripes.

therein). So, although taken seriously, quantum stripes have not been seen as a proven fact.

Reznik *et al.*¹ present a new indicator of quantum stripes by exploiting the motions of the ions that form the copper oxide lattice. These motions give rise to quantized lattice vibrations, known as phonons, that can be easily observed, again by inelastic neutron scattering. Electronic stripes must also undergo coherent vibrations, but these cannot be

seen directly. When the phonons and the stripe vibrations enter resonance, however, they are expected to interact strongly and cause a characteristic anomaly in the spectrum of the phonons. Reznik and colleagues observe just such a phenomenon in a copper oxide with static stripes.

The key is that these anomalies will occur on a timescale that is shorter than the quantum fluctuation time of the delocalized

quantum stripes. The quantum stripes will therefore seem to come to a standstill when viewed through the phonons, and the phonon anomaly should persist even when there is no sign of static stripes. This is exactly what Reznik *et al.* observe¹: even in the best superconductors, which show no sign of static stripes, the anomaly is blurred, but it is still clearly discernible.

So is this the turning point in the stripe wars? Although quantum stripes are more elusive than nuclear submarines — all signals of them have so far been indirect — the strength of the idea is that this single hypothesis explains a world of strange behaviours. Viewed from that angle, the defenders of the conventional BCS model might seem like medieval defenders of a geocentric cosmos, forced by observations to add ever more epicycles to their already baroque universe. On the other hand, there is no a priori need for electrons in crystals to behave in aesthetically pleasing ways; the epicycles could still be the truth. For one side or the other to win the war, a way of sending a sortie for direct reconnaissance of quantum stripes must be found.

Jan Zaanen is at the Instituut-Lorentz, Universiteit Leiden, PO Box 9506, 2300 RA Leiden, The Netherlands. e-mail: jan@lorentz.leidenuniv.nl

1. Reznik, D. *et al.* *Nature* **440**, 1170–1173 (2006).
2. *Nature Phys.* **2**, 138–143 (2006).
3. Tranquada, J. M. *et al.* *Nature* **375**, 561–563 (1995).
4. Zaanen, J., Nussinov, Z. & Mukhin, S. I. *Ann. Phys. (NY)* **310**, 181–260 (2004).
5. Kivelson, S. A. *et al.* *Rev. Mod. Phys.* **75**, 1201–1241 (2003).
6. Eremin, I. *et al.* *Phys. Rev. Lett.* **94**, 147001 (2005).

BIOMATERIALS

Gripping stuff

The aquatic bacterium *Caulobacter crescentus* can come, quite literally, to a sticky end. As Peter H. Tsang and colleagues report (*Proc. Natl Acad. Sci. USA* **103**, 5764–5768; 2006), the bacterium's long, tail-like anchor sticks it so tightly to a supporting surface that it often tears apart when a detaching force is applied, rather than relinquishing its grip. The adhesion is thought to be the strongest of biological origin yet discovered.

When *C. crescentus* (two are pictured here, spawning clones) is in its non-motile (stuck) state, its anchor — appropriately named its 'holdfast' — binds the bacterium to the surface, and a stalk connects the holdfast to the cell body. The authors used atomic force microscopy to record the force needed to detach a non-motile cell from a micropipette by means of a suction pipette

oriented perpendicular to the pulling direction. They measured the area of coverage of the holdfast and other dimensions of the bacterium, working out average geometries and finally applying a mathematical technique known as finite-element analysis to calculate the adhesion strength.

The holdfast enables the bacterium to remain stuck to the surface even in strong jets of water, and Tsang *et al.* calculate that, were it to cover an area of 1 cm², it could support a weight of 680 kg, even on a wet surface. That exceeds all other known cell-adhesion capabilities — including the sticking power of the much-studied gecko's foot. And because, owing to the geometry of *C. crescentus*, the adhesion often failed through fracture of the cell stalk rather than at the adhesion interface, the true adhesion strength



at the interface could be still higher.

Polymers of a sugar-based molecule called *N*-acetylglucosamine are known to be present in the bacterium's adhesive plaque. The authors found that when they treated the polymers with an enzyme to break them down, the strength of the

bacterial attachment was reduced.

However, the detailed physical and chemical mechanisms of *C. crescentus*'s adhesive abilities remain to be revealed. Their elucidation could trigger the development of a new range of synthetic adhesives.

Rosamund Daw

Y. BRUN

PALAEOCLIMATOLOGY

The woods fill up with snow

Michael N. Evans

Palaeoclimatological evidence covering the past millennium suggests that the global water cycle has changed in the past century. Agreement with climate models points to human activity as the main cause.

Two principal uncertainties in predicting climate change are the net effect of water in the climate system and the way in which water will be redistributed over the surface of the planet¹. On page 1179 of this issue, Treydte *et al.*² present a reconstruction of precipitation in central Asia for 826–1998 AD, developed from tree-ring data. Their results reveal a striking increase in snowfall over the past century compared with that over the previous millennium. This picture and other observations from around the world are concurring with predictions made by climate models, suggesting that the most pertinent aspects of the water cycle are adequately represented in the models. Furthermore, it seems that recent changes in precipitation patterns probably exceed the range of natural variability estimated for the past several hundred to one thousand years.

Resolving the influence of water on climate change is tricky, because the water cycle has interlinked effects on Earth's radiation budget, on the circulation of the atmosphere and ocean, and on the composition of any ecosystem. In turn, these processes combine to determine the distribution of moisture. The effects may occur on timescales from minutes to centuries, and on spatial scales from metres to hundreds of kilometres (Fig. 1)^{3–6}. Models incorporating the full richness of these processes and couplings could predict where, when and how climate change will manifest

itself — for instance, as a change in the frequency of devastating droughts. Confirming that certain climatic features have changed in concert with the rise of greenhouse-gas emissions during the past century would improve our confidence in such models.

In pursuit of this goal, Treydte *et al.*² developed a thousand-year estimate of the precipitation in central Asia — one of only a handful of well-dated, annually resolved reconstructions of climate variability on this timescale. Thanks to a series of laboratory, field and modelling studies⁷, the effects of environmental changes on the relative abundances of oxygen isotopes in the cellulose (cell-wall) component of wood are relatively well established (Fig. 2). About 70% of the variation in oxygen isotope composition of tree cellulose is determined by the isotopes in the source water taken up by the tree. The remainder is set by isotopic enrichment of water in the leaf or needles by evaporation before photosynthesis and the production of cellulose.

At high elevations in central Asia, precipitation is mostly snowfall. Treydte *et al.* use linear regression to demonstrate a link between the oxygen isotope composition in the cellulose of juniper trees in the region, and the average precipitation per water-year (from October to September) during the past century, observed at the handful of meteorological stations in the region. The authors show that greater

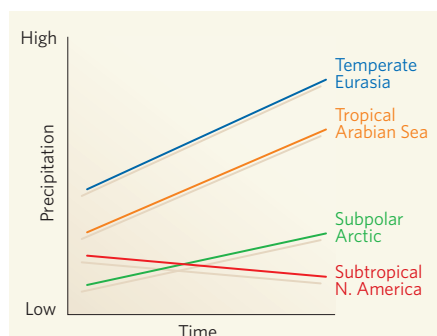


Figure 2 | Twentieth-century trends in palaeo-estimates of precipitation from around the world.

The palaeo-observations used to deduce variations in precipitation show large changes over the twentieth century compared with the past several centuries². However, they are qualitatively consistent with predictions of the general circulation model⁹ for regional changes in the global water cycle in response to increases in greenhouse-gas concentrations. The plots are not to scale and are meant to indicate qualitative direction of change only. Data sources: subtropical western North America, tree-ring widths; temperate Eurasia, tree-ring widths and oxygen isotopes; tropical Arabian Sea, ocean sediments; subpolar Arctic, observed and modelled tree-ring widths¹⁰.

precipitation is associated with lower concentrations of the oxygen isotope ¹⁸O. This is qualitatively consistent with the expected isotopic effects, and with observed global patterns in the stable isotopic composition of precipitation⁸.

The authors use the relationship between snowfall and oxygen isotope composition to reconstruct a time series of precipitation from tree-ring data. The data set has a high standard of intra- and inter-site replication that is rarely matched even in modern high-resolution isotopic palaeoclimate studies. Five records from one site share, on average, 36% of their variances; three twentieth-century records from three nearby sites share, on average, 25% of their variances. The time series of more than 1,100 years is composed of overlapping segments of pooled samples from at least 14 cores from seven trees for each year of record. It shows the same statistical relationship and quality of reconstruction as do the twentieth-century time series used by the authors to demonstrate the snowfall–isotope link. For palaeoclimate data from tree rings, these are very convincing results, and they indicate a common, interpretable signal of climate change.

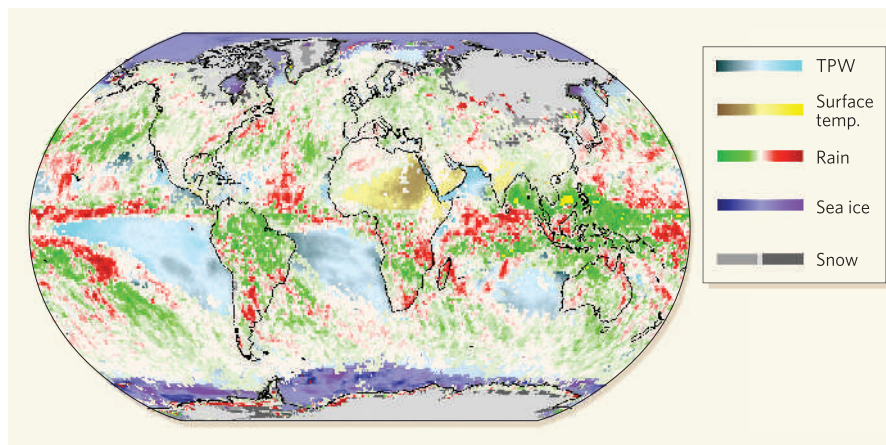


Figure 1 | Distribution of water in the climate system. Measurements are from the NOAA-15 Advanced Microwave Sounding Unit (AMSU) instrument⁶. The map shows the difference between December 2004, a weak El Niño/Southern Oscillation (ENSO) warm-phase event, and December 2005, a weak ENSO cold-phase event. ENSO events result in global redistributions of fresh water, which amplifies local, regional and global climate variability. Differences in total atmospheric precipitable water (TPW) range from −10 to +10 mm; in rainfall range from −1.0 to +1.0 mm per hour; in sea-ice concentration range from −50 to +50%; in snow cover range from −0.5 to +0.5 fraction of month for which snow cover was found in the grid box; and in surface temperature range from −10 to +10 K. (Courtesy of H. Meng and R. R. Ferraro, NOAA/NESDIS.)

In Treydte and colleagues' reconstruction, the greatest sustained increase in precipitation occurred in the past 150 years, which is roughly coincident with the Industrial Revolution and greenhouse-gas increases. As in the poem by Robert Frost, the woods have been filling up with snow.

Does this agree with expectations based on model-derived climate-forecasting simulations that include realistic depictions of the effects of greenhouse-gas emissions? Roughly, yes. The authors review the relevant literature and find that global precipitation increased by 1% per decade during the twentieth century. Furthermore, this and other palaeoclimate records (Fig. 2) suggest not only that the observed precipitation trends over the past century are unusual in the context of previous centuries, but also that observed drying in the subtropics and moistening in the tropics and high latitudes are consistent with model predictions⁹.

The accuracy of climate-model predictions for trends in precipitation depends on whether the many interlinked roles of water in climate are correctly depicted. On the other hand, many uncertainties remain in interpreting long-term features in the palaeoclimate records. So agreement between the climate model and the estimates of rainfall deduced from various measures is compelling. A denser network of high-quality gauges of precipitation, created from a broad range of locations and biogeochemical data sources, will permit palaeoclimatologists to distinguish more clearly between the natural variability of the water cycle and alterations induced by greenhouse warming. This will in turn allow the uncertainties in data archives to be evaluated. It will also allow better definition of the spatial patterns in climate features during the past few centuries to millennia, and testing of climate-change hypotheses developed from simulations. All of these advances will assist the crucial work of climate-change detection and attribution¹.

Michael N. Evans is in the Laboratory of Tree-Ring Research, University of Arizona, 105 West Stadium, Tucson, Arizona 85721, USA.
e-mail: mevans@ltrr.arizona.edu

MOLECULAR BIOLOGY

Singled out for integration

Michael Chandler

Bacteria regularly swap genes among themselves. The structure of an enzyme that mediates the spread of certain DNA elements reveals the unusual way in which they take up residence in new genomes.

The transfer of genes among bacteria is crucial in generating diversity, allowing bacteria to adapt quickly to changing environments. Mobile DNA elements called integron cassettes are major vehicles in such gene exchange. In this issue, MacDonald *et al.* (page 1157)¹ report the structure of an enzyme that inserts these cassettes into genomes, providing some much-needed insight into how these ubiquitous genetic building-blocks move among bacteria.

Integron cassettes each contain a single gene from among a large range of gene types, including those encoding antibiotic resistance factors, virulence factors and restriction endonuclease enzymes². These cassettes are inserted into DNA modules called integrons (Fig. 1a), present in many genomes. The integrons supply both the enzyme — an 'integrase' — necessary to insert the cassettes, and the signals required for expression of the inserted genes. Integrons accumulate integron cassettes into tandem arrays, which can contain several hundred individual units each carrying a different gene³. Given the biological impact of integrons, molecular biologists are keen to understand how, at the molecular level, the cassettes are captured and assembled into the integron.

Earlier studies had identified circular forms of integron cassettes that are believed to be intermediates in their movement between genomes³. These circular cassettes were proposed to be captured by an integron through recombination — a site-specific integration reaction — between a short sequence in the integron (called *attI*) and another sequence in the circular cassette (*attC*)⁴ (Fig. 1).

But how are the circular cassettes made? Part of the answer is provided by MacDonald *et al.*¹, who present a crystal structure of an integron-cassette recombination site bound by the integrase IntI. IntI is a member of a large family of related enzymes that recognize and recombine double-stranded DNA using a tyrosine residue to break and rejoin both strands⁵. But IntI does not bind to double-stranded *attC* DNA. Instead, it binds specifically to a unique single strand of the integron cassette^{6–8}, suggesting that the cassette might be excised from the integron as a single-stranded rather than a double-stranded DNA circle.

The structure shows a complex of four IntI molecules bound to two *attC* sites. Strikingly, each *att* site is single stranded, but folded to

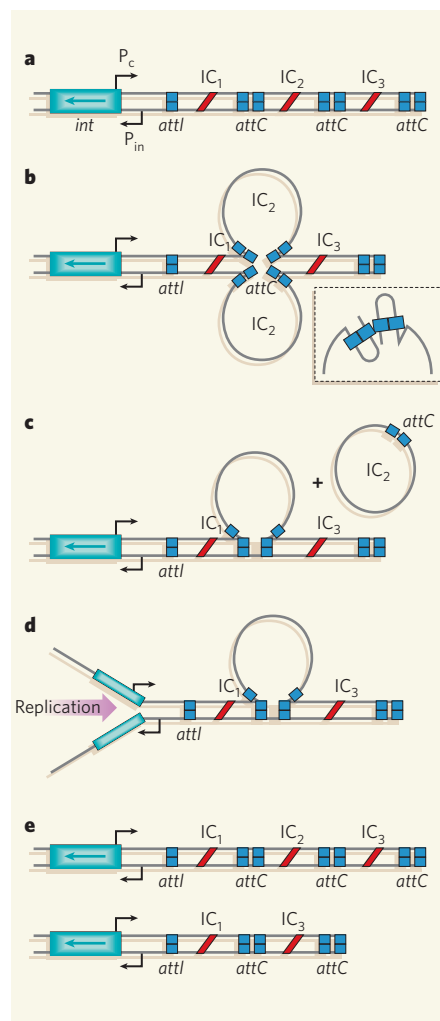


Figure 1 | The integron and integron cassette excision. **a**, A typical integron containing the gene (*int*) encoding the IntI enzyme and the regulatory region (P_{in}) that drives *int* gene expression. The integron also carries the regulatory region (P_c) that controls expression of the genes in an array of integron cassettes (IC_1 , IC_2 , IC_3). The *att* sequences (blue boxes) are necessary for cassette insertion and excision. **b**, Extrusion of the single-stranded cassette during cassette excision and folding of the unique (bottom) strand to generate two functional *attC* sites (inset). **c**, Recombination between the *attC* sites generates the single-stranded cassette, which can then be captured by a second integron. This leaves the original donor integron with a single *attC* joining one strand and a single-strand loop. **d,e**, Replication generates two daughter molecules: one with the cassette deleted and the other retaining a full copy.

1. Intergovernmental Panel of Climate Change *Climate Change 2001: Synthesis Report* (eds Watson, R. T. *et al.*) (Cambridge Univ. Press, 2001).
2. Treydte, K. S. *et al.* *Nature* **440**, 1179–1182 (2006).
3. Hartmann, D. L. *Global Physical Climatology* (Int. Geophys. Ser. Vol. 56, Academic, New York, 1994).
4. Gill, A. E. *Atmosphere–Ocean Dynamics* (Int. Geophys. Ser. Vol. 30, Academic, New York, 1982).
5. Bonan, G. B., Levis, S., Sitch, S., Vertenstein, M. & Oleson, K. W. *Glob. Change Biol.* **9**, 1543–1566 (2003).
6. Ferraro, R. R. *et al.* *IEEE Trans. Geosci. Remote Sensing* **43**, 1036–1049 (2005).
7. Roden, J. S., Lin, G. & Ehleringer, J. R. *Geochim. Cosmochim. Acta* **64**, 21–35 (2000).
8. Bowen, G. J. & Wilkenson, B. *Geology* **30**, 315–318 (2001).
9. Emori, S. & Brown, S. J. *Geophys. Res. Lett.* **32**, L17706 (2005).
10. Vaganov, E. A., Hughes, M. K., Kirdyanov, A. V., Schweingruber, F. H. & Silkin, P. P. *Nature* **400**, 149–151 (1999).

form a double-stranded site (Figs 1b, inset, and 2). Recombination between the folded *attC* sites could generate single-stranded cassette circles (Fig. 1b) that might then be captured by a second integron.

MacDonald and colleagues' work also gives an invaluable view into how integron cassette recombination is regulated. Recombination occurs by exchange of DNA strands between two DNA partners that must first be aligned at their *att* sites. The *att* sequences recognized by most Int family enzymes have two adjacent protein-binding sites, left (L) and right (R), separated by a small sequence within which the exchange of DNA strands occurs^{4,5}. Alignment of the recombination partners involves interactions between four Int molecules, with two bound to each *att* — forming the so-called synaptic complex (Fig. 2a). The distinguishing feature of the integron cassette *attC* sequence is that it carries two imperfect L–R pairs that, individually, are inactive. However, by folding from a single strand, they form a single functional site (Fig. 1b) in which R is a fully paired DNA helix, and L and its flanking sequence carry several extrahelical bases (Fig. 2a). Recognition of these bases proves to be the key to a surprising regulation mechanism.

To ensure that only the correct single strand is cleaved and rejoined to form the single-stranded cassette circle, two *att* subunits in the synaptic complex must be maintained in an inactive configuration while the other two are active. It is recognition of the extrahelical bases that correctly positions IntI, resulting in changes in the enzyme configuration that inactivate one pair of subunits while activating the other. An extrahelical thymine residue in the L-site flank blocks L-bound subunits in an inactive configuration where the catalytic tyrosine of IntI is too distant from its target bond to catalyse DNA-strand cleavage. R-bound subunits can then reach across the synapse and 'grasp' the opposite DNA recombination partner, establishing contacts with an extrahelical guanine residue in its L site. This stabilizes the complex, permits additional cross contacts, and brings the catalytic tyrosine close to its target bond for DNA cleavage (Fig. 2b).

The structure reported by MacDonald *et al.* includes two *attC* sites. Although *attC*–*attC* recombination must occur during cassette excision, cassette integration would involve pairing between the folded *attC* and the integron *attI* site⁶. When bound by IntI, *attI* is in its double-stranded form. Excisive and integrative synapses are thus architecturally different. To understand fully the capture and transmission of integron cassettes, future studies will need to determine the structure of the integrative complex.

Although unusual, the type of single-stranded DNA recombination hinted at by the IntI structure is not limited to integron cassettes. A similar mechanism has been

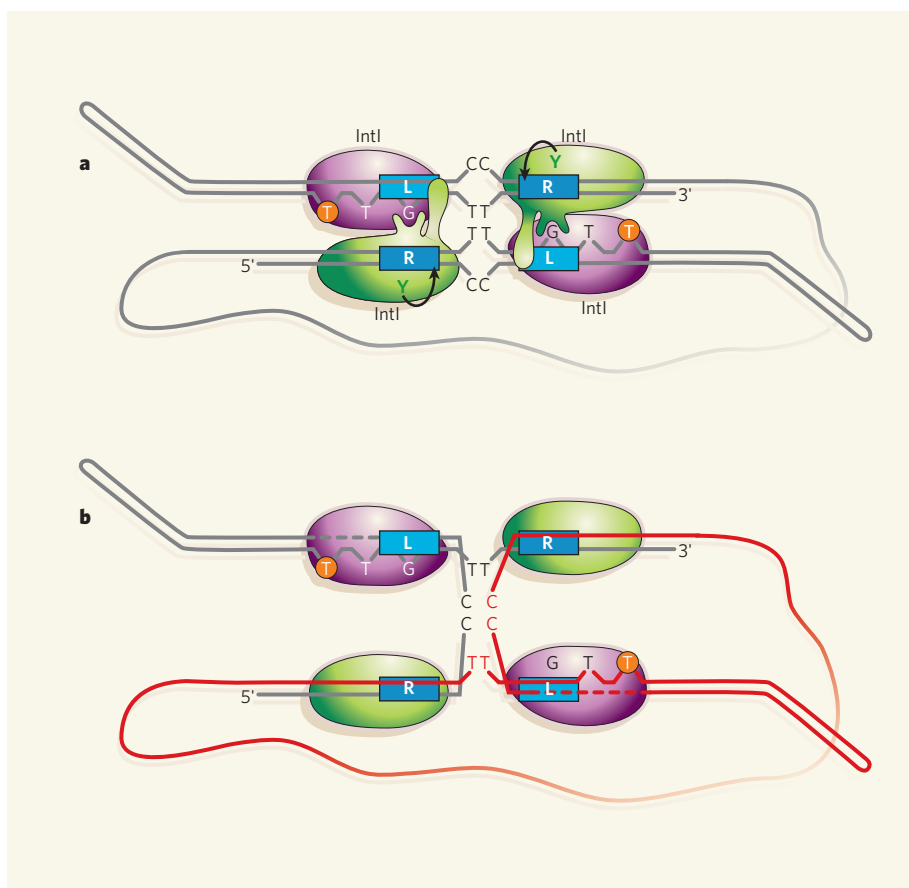


Figure 2 | Structure of the recombination complex reported by MacDonald *et al.*¹ **a**, The two folded *attC* sites flanking an integron cassette are aligned and bridged by four IntI molecules. Only the single DNA strand of the cassette that will form a circle is shown (grey). The left (L) and right (R) IntI-binding sites are boxed and the extrahelical bases are shown. The inactive IntI monomers (purple) bound to L are inactivated by conformational changes resulting from recognition of an extrahelical thymine (T) residue (orange circle). The active monomers (green), bound to R, reach across and grasp the extrahelical guanine (G) from the opposing L site. This results in a conformational change allowing a neighbouring part of the protein to generate further cross contacts. **b**, Cleavage and rejoining promoted by the active monomers results in formation of a single-stranded circular cassette (red). For simplicity, the partner single strand is not shown.

proposed for the integration of a bacterial virus, bacteriophage CTX ϕ , into the *Vibrio cholerae* genome⁹. Moreover, related models would also explain other systems involved in horizontal gene transfer, including plasmid and viral replication (REP proteins), conjugal plasmid transfer from cell to cell (relaxases) and certain transposase enzymes^{10–14}.

The intriguing results of MacDonald *et al.* raise an obvious but significant question: how are the single-stranded recombination substrates generated in nature? Candidate mechanisms would be: during DNA replication; through natural genetic transformation, when the bacteria take up DNA from their environment; or during conjugal transfer when carried by a transmissible plasmid. These processes can all generate single-stranded DNA. Gene transfer by transformation or conjugation with cells carrying a target integron would provide an excellent mechanism for activating integron cassette excision and therefore for promoting dispersion of cassettes.

Michael Chandler is at the Laboratoire de

Microbiologie et Génétique Moléculaire (CNRS, UMR5100), Campus Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France. e-mail: mike@ibcg.biotoul.fr

- MacDonald, D., Demarre, G., Bouvier, M., Mazel, D. & Gopaul, D. N. *Nature* **440**, 1157–1162 (2006).
- Rowe-Magnus, D. A., Guerout, A. M., Biskri, L., Bouige, P. & Mazel, D. *Genome Res.* **13**, 428–442 (2003).
- Collis, C. M. & Hall, R. M. *Mol. Microbiol.* **6**, 2875–2885 (1992).
- Recchia, G. D. & Sherratt, D. J. in *Mobile DNA II* (eds Craig, N., Craigie, R., Gellert, M. & Lambowitz, A. M.) 162–176 (ASM, Washington DC, 2002).
- Van Duyn, G. D. in *Mobile DNA II* (eds Craig, N., Craigie, R., Gellert, M. & Lambowitz, A. M.) 93–117 (ASM, Washington DC, 2002).
- Bouvier, M., Demarre, G. & Mazel, D. *EMBO J.* **24**, 4356–4367 (2005).
- Francia, M. V., Zabala, J. C., de la Cruz, F. & Garcia Lobo, J. M. *J. Bacteriol.* **181**, 6844–6849 (1999).
- Johansson, C., Kamali-Moghadam, M. & Sundstrom, L. *Nucleic Acids Res.* **32**, 4033–4043 (2004).
- Val, M. E. *et al.* *Mol. Cell* **19**, 433–435 (2005).
- Hickman, A. B., Ronning, D. R., Perez, Z. N., Kotin, R. M. & Dyda, F. *Mol. Cell* **13**, 403–414 (2004).
- Guasch, A. *et al.* *Nature Struct. Biol.* **10**, 1002–1010 (2003).
- Larkin, C. *et al.* *Structure* **13**, 1533–1544 (2005).
- Ton-Hoang, B. *et al.* *EMBO J.* **24**, 3325–3338 (2005).
- Ronning, D. R. *et al.* *Mol. Cell* **20**, 143–154 (2005).

BRIEF COMMUNICATIONS

Therapeutic gene causing lymphoma

Insight into risks posed by corrective gene therapy comes from an immunodeficient mouse model.

The development of T-cell leukaemia following the otherwise successful treatment of three patients with X-linked severe combined immune deficiency (X-SCID) in gene-therapy trials using haematopoietic stem cells¹ has led to a re-evaluation of this approach². Using a mouse model for gene therapy of X-SCID, we find that the corrective therapeutic gene *IL2RG* itself can act as a contributor to the genesis of T-cell lymphomas, with one-third of animals being affected. Gene-therapy trials for X-SCID, which have been based on the assumption that *IL2RG* is minimally oncogenic^{3–7}, may therefore pose some risk to patients.

The syndrome X-SCID is caused by faulty expression of the γ -chain of the interleukin-2 receptor (IL2RG), which results in diminished lymphoid-cell survival and proliferation. Gene therapy can restore IL2RG expression and hence adaptive immunity in X-SCID patients. The development of leukaemia after this gene therapy has been attributed to the upregulated expression of the oncogene *LMO2* as a result of vector integration².

To investigate the origin of these adverse events, we expressed *IL2RG* inserted into a lentiviral vector (LV) in a murine model of X-SCID, and followed the fates of mice for up to 1.5 years post-transplantation. Unexpectedly, 33% of these mice ($n = 15$) developed T-cell lymphomas that were associated with a gross thymic mass (Fig. 1, and see supplementary information). Lymphomic tissues shared a common lymphomic stem cell, with similar vector-integration sites being evident in the DNA of the thymus, bone marrow and spleen of individual mice; however, no common integration targets were found between mice.

As expected, T-cell lymphomas were also detected in positive-control mice transduced with *LMO2* in a lentiviral vector ($n = 12$; prevalence, 50%; the cell phenotype has been described⁸), although none of the mock-transduced mice ($n = 15$) or control-vector-transduced mice (LV-GFP; $n = 15$; Fig. 1) developed thymomas. These results, generated from independent transplant experiments with 76 mice ranging from 41 to 81 weeks post-transplant, indicate that insertional mutagenesis was not

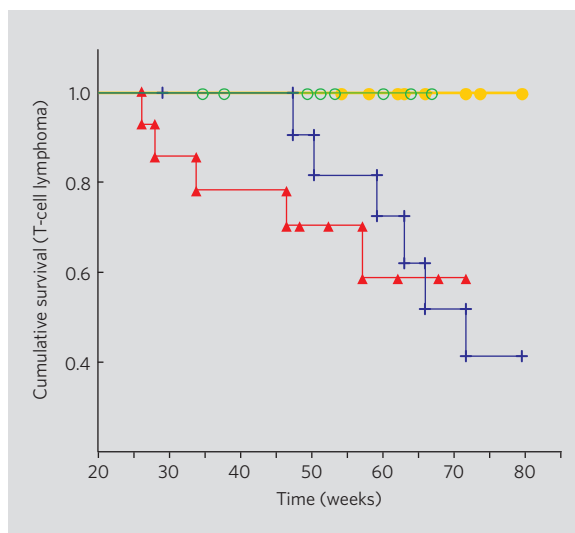


Figure 1 Kaplan-Meier statistical analysis showing occurrence in mice of T-cell lymphomas. Significantly more T-cell lymphomas developed in animals transplanted with cells transduced with LV-IL2RG (triangles) compared with controls that were mock-transduced or LV-GFP vector-transduced (filled circles and open circles, respectively). LV-GFP expresses the marker gene for green fluorescence protein and is used to verify that the principal lymphomagenic event is neither insertional mutagenesis nor the presence of oncogenic elements within the vector backbone. Crosses indicate positive controls transplanted with LV-*LMO2* containing the T-cell oncogene *LMO2*. We found 33% of LV-IL2RG mice died within 81 weeks of transplantation with either X-SCID or wild-type *IL2RG* (CD45.1) haematopoietic donor cells.

the principal cause of lymphoma in our mice.

How has the oncogenicity of *IL2RG* been overlooked? Worldwide, some 88 mice have been treated with *IL2RG* in retroviral vectors^{3–5,9–11}, but these studies were limited by their duration, which usually did not exceed 6 months post-transplant. (Our first *IL2RG*-induced lymphoma appeared 6 months post-transplant.) Longer-term analysis did not reveal leukaemogenesis in dogs, rhesus macaques or sheep carrying human chimeric genes for up to one year post-transplant, but this may have been attributable to features inherent in large-animal models^{6,7}. In the human gene-therapy trials¹, leukaemias did not appear until 2–3 years after treatment.

Lymphomagenesis in our mice may result from altered signalling through one or more of the interleukin receptors that contain the IL2RG subunit (including those for interleukins 2, 4, 7, 9, 15 and 21, most of which have been associated with T-cell neoplasias). Altered cell-survival and/or tumour-suppressor

functions due to abnormal receptor expression are known to occur in B-cell-receptor signalling, where overexpression of Btk blocks B-cell development but underexpression leads to B-cell neoplasia¹². The long latency period before lymphomas develop in our LV-IL2RG-treated mice indicates that other complementary mutations may be required for lymphomagenesis.

Our results indicate that preclinical experimental treatments involving transgenes should include long-term follow-up before they enter clinical trials. Moreover, our findings highlight the need for continued development of vectors capable of regulated therapeutic gene expression.

Niels-Bjarne Woods*, Virginie Bottero*, Manfred Schmidt†‡, Christof von Kalle†, Inder M. Verma*

*Laboratory of Genetics, The Salk Institute for Biological Studies, La Jolla, California 92037, USA

e-mail: verma@salk.edu

†National Center for Tumor Diseases

Heidelberg and Department of Translational Oncology, German Cancer Research Center, Otto-Meyerhof-Zentrum Im Neuenheimer Feld 350, 69120 Heidelberg, Germany

‡Department of Internal Medicine I, University Hospital, 79106 Freiburg, and Institute of Molecular Medicine and Cell Research, Albert-Ludwigs-University, 79104 Freiburg, Germany

1. Hacein-Bey-Abina, S. et al. *N. Engl. J. Med.* **346**, 1185–1193 (2002).
2. Hacein-Bey-Abina, S. et al. *Science* **302**, 415–419 (2003).
3. Soudais, C. et al. *Blood* **95**, 3071–3077 (2000).
4. Otsu, M., Sugamura, K. & Candotti, F. *Blood* **97**, 1618–1624 (2001).
5. Aviles Mendoza, G. J. et al. *Mol. Ther.* **3**, 565–573 (2001).
6. Whitwam, T. et al. *Blood* **92**, 1565–1575 (1998).
7. An, D. S. et al. *J. Virol.* **75**, 3547–3555 (2001).
8. Larson, R. C., Osada, H., Larson, T. A., Lavenir, I. & Rabbitts, T. H. *Oncogene* **11**, 853–862 (1995).
9. Lo, M. et al. *Blood* **94**, 3027–3036 (1999).
10. Otsu, M. et al. *Mol. Ther.* **1**, 145–153 (2000).
11. Otsu, M., Sugamura, K. & Candotti, F. *Hum. Gene Ther.* **11**, 2051–2056 (2000).
12. Kersseboom, R. et al. *J. Exp. Med.* **198**, 91–98 (2003).

Supplementary information accompanies this communication on Nature's website.

Received 18 October 2005; accepted 28 March 2006.

Competing financial interests: declared none.

doi:10.1038/4401123a

BRIEF COMMUNICATIONS ARISING online
www.nature.com/bca see Nature contents.

MOLECULAR VIROLOGY

Was the 1918 pandemic caused by a bird flu?

Arising from: J. K. Taubenberger *et al.* *Nature* **437**, 889–893 (2005)

Taubenberger *et al.*¹ have sequenced the polymerase genes of the pandemic 'Spanish' influenza A virus of 1918, thereby completing the decoding of the genome of this virus^{2–6}. The authors conclude from these sequences that the virus jumped from birds to humans shortly before the start of the pandemic and that it was not derived from earlier viruses by gene shuffling, a process called reassortment. However, we believe that their evidence does not convincingly support these conclusions and that some of their results even indicate that, on the contrary, the virus evolved in mammals before the pandemic began and that it was a reassortant. In light of this alternative interpretation, we suggest that the current intense surveillance of influenza viruses should be broadened to include mammalian sources.

The inferences by Taubenberger *et al.*¹ are based on two lines of evidence. First, they report phylogenetic trees for nine of the eleven genes of the virus. The trees would have supported the conclusions if, in all or at least in most of them, the 1918 virus linked the bird influenza viruses with the later human and swine influenza viruses: that is, if it had been placed on the root of the cluster of mammalian viruses (Fig. 1a). That topology would indicate that the 1918 virus was the ancestor of all post-1918 human influenza viruses, as Taubenberger *et al.*¹ assume.

However, none of the nine phylogenetic trees described by Taubenberger and his co-workers, including those of the polymerases, shows that topology^{1–6}. In five of the published trees, the 1918 gene is placed next to the main cluster of human influenza viruses, and the classical swine influenzas are linked to the tree between the branches of the human and avian viruses (Fig. 1b). The other four trees have the reverse topology; the 1918 gene lies next to the classical swine viruses, and the main cluster of human viruses is between them and the avian viruses (Fig. 1c). The combined mammalian cluster is linked to the bird cluster directly in five of the trees, but in the other four trees it is linked first to other lineages from mammals (that is, those of pigs and horses).

Taubenberger and colleagues^{1–6} report that the 1918 virus was placed with the bird influenza viruses in some unpublished trees and that in all the trees, both published and unpublished, the 1918 virus lies close to the root of the mammalian lineages of influenza viruses. However, the 1918 virus does not lie on the root, and therefore the phylogenetic analyses of Taubenberger *et al.* do not support their conclusions: their results indicate instead

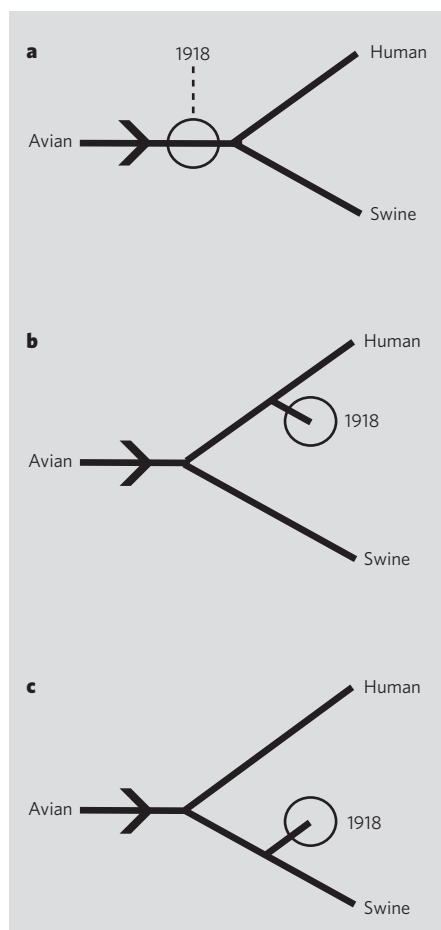


Figure 1 | Hypothetical and observed phylogeny of influenza-virus genes. **a**, Phylogenetic tree that would be expected if the 1918 virus had come directly from birds. The avian influenza-virus source would be linked to the later influenzas isolated from mammals, namely on the branch linking the avian viruses to the mammalian viruses. **b**, Summary of the trees found by Taubenberger and co-workers^{1–6} for five of the influenza-virus genes (*PA*, *PB1*, *HA*, *M1*, *M2*; see ref. 1 for details), which placed the 1918 virus next to the later human influenzas: that is, they are sister groups. **c**, Summary of the remaining four influenza virus gene trees found by Taubenberger and co-workers^{1–6} (*PB2*, *NP*, *NS1*, *NA*; see ref. 1 for details), which placed the 1918 virus next to the later swine viruses. The avian-virus source cluster in most of the gene trees also includes virus isolates from mammals.

that the virus evolved in people or pigs for an unknown period of time before the pandemic started and that the virus may have been a reassortant.

The second line of evidence used by Taubenberger *et al.*¹ to support their conclusions comes from comparing similarities in

polymerase amino-acid sequences. Phylogenetic analysis is usually preferred to this approach for studying origins, because similarities can sometimes occur by coincidence through parallel evolution. Indeed, Taubenberger *et al.*¹ discovered examples of identical amino-acid residues in bird and human influenza virus polymerases that resulted from parallel evolution, and we contend that those discoveries undermine their conclusions.

Although the 1918 virus polymerase proteins are placed closest to the typical sequences of some bird influenza viruses, the ranking depends on very few residues; parallel evolution could therefore have affected the ranking. In contrast, when nucleotide similarity was measured, rather than amino-acid similarity, the authors found the 1918 virus to be significantly closer to influenza viruses from mammals. The evidence of sequence similarity uncovered by Taubenberger *et al.*¹ is important for understanding how influenza viruses adapt to humans, but it does not prove their conclusions.

The events that led to the emergence of the 1918 virus are unclear and will probably remain so, at least until the immediate ancestors of the virus have been characterized and the mutation rate of the virus lineage is known. Data from other influenzas show that the mutation rate varies from close to 0% to almost 1% per year⁷, introducing uncertainty about timing inferred from influenza-virus sequences. The haemagglutinin gene from an influenza virus from a goose captured in 1917 has already been sequenced⁸: perhaps unluckily, the 1917 virus was not an ancestor of the 1918 pandemic virus, but it was similar to contemporary H1N1 avian influenzas.

Mark J. Gibbs*, **Adrian J. Gibbs†**

*School of Botany and Zoology, The Australian National University, Canberra, ACT 0200, Australia

e-mail: mark.gibbs@anu.edu.au

†7 Hutt Street, Yarralumla, ACT 2600, Australia

1. Taubenberger, J. K. *et al.* *Nature* **437**, 889–893 (2005).
2. Basler, C. F. *et al.* *Proc. Natl Acad. Sci. USA* **98**, 2746–2751 (2001).
3. Reid, A. H., Fanning, T. G., Hultin, J. V. & Taubenberger, J. K. *Proc. Natl Acad. Sci. USA* **96**, 1651–1656 (1999).
4. Reid, A. H., Fanning, T. G., Janczewski, T. A. & Taubenberger, J. K. *Proc. Natl Acad. Sci. USA* **97**, 6785–6790 (2000).
5. Reid, A. H., Fanning, T. G., Janczewski, T. A., McCall, S. & Taubenberger, J. K. *J. Virol.* **76**, 10717–10723 (2002).
6. Reid, A. H., Fanning, T. G., Janczewski, T. A., Lourens, R. M. & Taubenberger, J. K. *J. Virol.* **78**, 12462–12470 (2004).
7. Webster, R. G., Bean, W. G., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. *Microbiol. Rev.* **56**, 152–179 (1992).
8. Fanning, T. G. *et al.* *J. Virol.* **76**, 7860–7862 (2002).

doi:10.1038/nature04823

MOLECULAR VIROLOGY

Was the 1918 flu avian in origin?

Arising from: J. K. Taubenberger *et al.* *Nature* **437**, 889–893 (2005)

Taubenberger *et al.*¹ claim that the 1918 influenza virus was derived from an avian source and adapted to humans shortly before the pandemic. However, we do not believe that this conclusion, which has been widely disseminated in the popular press and in scientific journals^{2–4}, is supported by their phylogenetic evidence.

The authors' conclusion is based on DNA sequences from three RNA-polymerase genes, each of which resides on a separate RNA segment of the viral genome. The nucleotide phylogenies that they present¹ show, with strong statistical support, that the 1918 influenza virus is found within clades containing human

and classical swine influenza viruses and is not basal to those clades. Moreover, the relationship of the 1918 strain to avian strains (rather than to equine or other mammalian strains) is unresolved because the trees are unrooted.

The phylogenies described by Taubenberger *et al.*¹ contradict their main conclusions and are presented without discussion of the evolutionary relationships they imply. Instead, evolutionary conclusions are improperly drawn from a similarity between the 1918 and avian influenza viruses in the patterns of base-pair substitution (that is, the synonymous/non-synonymous and transition/transversion ratios, and variation in fourfold-degenerate

sites) and without consideration of the relative similarity of the 1918 flu in these traits to other mammalian strains. These are inappropriate characters from which to infer similarity by descent, and the data from the DNA sequences should not be discounted as they are superior indicators of phylogenetic relatedness.

Taubenberger *et al.*¹ also claim that the amino-acid sequences encoded by these RNA-polymerase genes support the avian nature of the 1918 virus. We reconstructed phylogenies using their amino-acid sequences⁵ and found that the 1918 virus falls within, and not basal to, clades containing strains from other mammalian hosts (Fig. 1). Sequences from the nucleoprotein gene of the influenza virus also fail to provide evidence that the 1918 strain is derived directly from an avian source⁵.

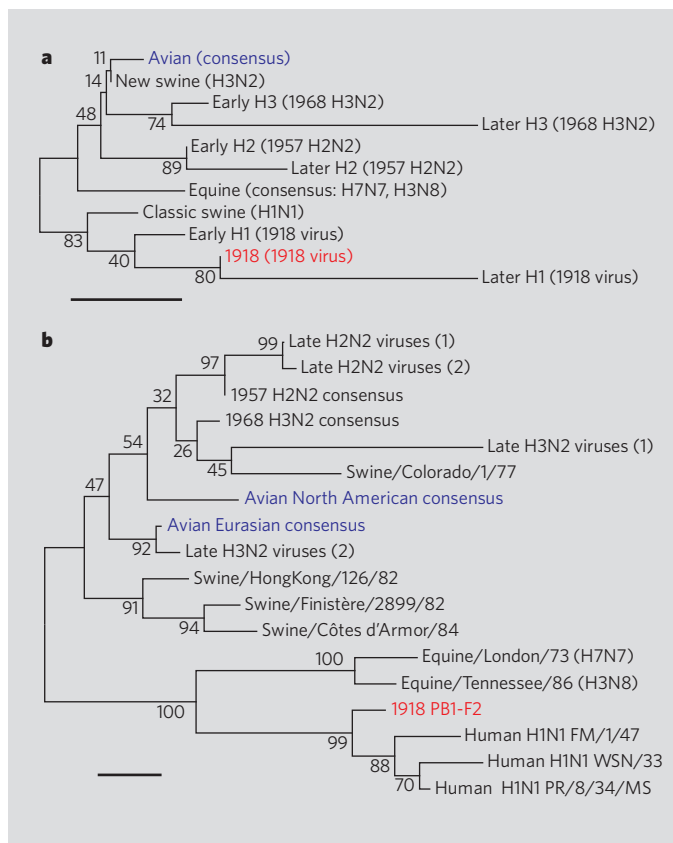
In support of their conclusions, Taubenberger *et al.*¹ cite prior phylogenetic studies⁶ indicating that the 1918 flu may have been derived from an avian source. However, those results simply show that the 1918 flu is phylogenetically unresolved and genetically equidistant from the North American and Asian clades of avian flu, which does not indicate emergence from an avian source.

By stating that the high pathogenicity of the 1918 virus is related to its emergence as a human-adapted avian influenza virus, the authors raise the possibility that an emerging avian strain could resemble the 1918 flu. This alarming implication, which is based on misinterpretation of the phylogenetic data, is completely unjustified and could seriously distort the public perception of disease risk, with grave economic and social consequences.

Janis Antonovics, Michael E. Hood, Christi Howell Baker

Department of Biology, University of Virginia, Charlottesville, Virginia 22904, USA
e-mail: ja8n@virginia.edu

Figure 1 | Neighbour-joining analysis of the PB1 subunit of the influenza virus A polymerase heterotrimer and of the peptide PB1-F2. **a, b,** Amino-acid alignments for **a**, PB1, and **b**, PB1-F2 (incomplete sequences omitted), taken from Supplementary Figure 2a, b of Taubenberger *et al.*¹. Data were analysed using MEGA 3.1 software⁷; bootstrap values represent 100 replications. Sample names are given in ref. 1. The 1918 and avian viruses are shown in red and blue text, respectively.



1. Taubenberger, J. K. *et al.* *Nature* **437**, 889–893 (2005).
2. Belshe, R. B. *N. Engl. J. Med.* **353**, 2209–2211 (2005).
3. Russell, C. J. & Webster, R. G. *Cell* **123**, 368–371 (2005).
4. Sharp, P. A. *Science* **310**, 17 (2005).
5. Reid, A. H., Fanning, T. G., Janczewski, T. A., Lourens, R. M. & Taubenberger, J. K. *J. Virol.* **78**, 12462–12470 (2004).
6. Reid, A. H., Taubenberger, J. K. & Fanning, T. G. *Nature Rev. Microbiol.* **2**, 909–914 (2004).
7. Kumar, S., Tamura, K. & Nei, M. *Brief. Bioinform.* **5**, 150–163 (2004).

doi:10.1038/nature04824

MOLECULAR VIROLOGY

Taubenberger *et al.* reply

Replying to: M. J. Gibbs & A. J. Gibbs *Nature* **440**, doi:10.1038/nature04823; J. Antonovics, M. E. Hood & C. H. Baker *Nature* **440**, doi:10.1038/nature04824 (2006)

We have proposed that the virus responsible for the 1918 human influenza pandemic was avian-like, which to us is a reasonable interpretation of all the available data from phylogenetic, sequence, serological and historical

analysis, combined with what is known of influenza A virus biology^{1–4}. However, Gibbs and Gibbs⁵ and Antonovics *et al.*⁶ question our inferred origin of the 1918 influenza virus.

Gibbs and Gibbs contest that the 1918 virus

evolved in mammals before the pandemic and was a reassortment⁵. They present three simplified trees showing the 1918 virus on the root of the mammalian clade, or in either the human or classical swine subclades, claiming that the 1918 virus was not placed on the root of the tree as expected. Likewise, Antonovics *et al.* state that the 1918 virus sequences are placed in the human and classical swine influenza clades and are not basal to them⁶. That is indeed true for the polymerase-gene trees shown, so here we assess this apparent discrepancy.

When branch lengths are examined, it is evident that the sequence of the 1918 virus is the closest to the common ancestor of the mammalian clade in the genes for each of the three polymerases (PA, PB1, PB2). Similarly, the haemagglutinin and neuraminidase genes of the 1918 virus are clearly placed very near the root of the mammalian clade^{7,8}. In contrast to results shown in Fig. 1 of Antonovics *et al.*⁶, which is derived from a subset of our data¹, a more extensive analysis using a larger data set of amino-acid sequences¹ places the 1918 virus closest to the root of the mammalian clade for each polymerase gene (for example, see Fig. 1).

In an earlier phylogenetic study of swine influenza viruses⁹, it was concluded that the internal genes of classical swine viruses are 'avian-like' and that these genes evolved concomitantly; furthermore, the results were in agreement with the historical account that the 1918 human influenza virus entered the pig population in 1918. We agree with the generally held conclusion that the human and classical swine H1N1 influenza viral lineages descended from the 1918 virus. It is important to note, however, that not all swine influenza viruses are from this lineage. The trees shown by Antonovics *et al.* show sequences from several swine influenza viruses that are not descended from the classical swine lineage, but rather from the European 'avian-like' swine influenza H1N1 lineages of the late 1970s¹⁰ or from an Asian H3N2 swine virus containing an avian-derived PB1 gene¹¹. These lineages share no evolutionary relationship with the 1918 virus.

We have never maintained that the virus entered the human population in 1918: rather, as described earlier^{2,12}, our claim that it entered the human population "shortly" before the pandemic should be interpreted as 'at least several years before the pandemic', as stated in our discussion¹. The path that the precursors of the 1918 pandemic virus took before emerging in humans in 1918 remains unknown. Phylogenetic analysis on its own cannot definitively resolve the issue. As in previous analyses⁹, we analysed the sequences of these genes for clues about their origins and found that the proteins encoded by the 1918 polymerase genes were avian-like in all cases.

Other scenarios can be derived from the available data to explain the origin of the 1918 human influenza virus. For example, a virus from an unknown avian-like source could have infected an unknown mammal, where it may have evolved for several years before

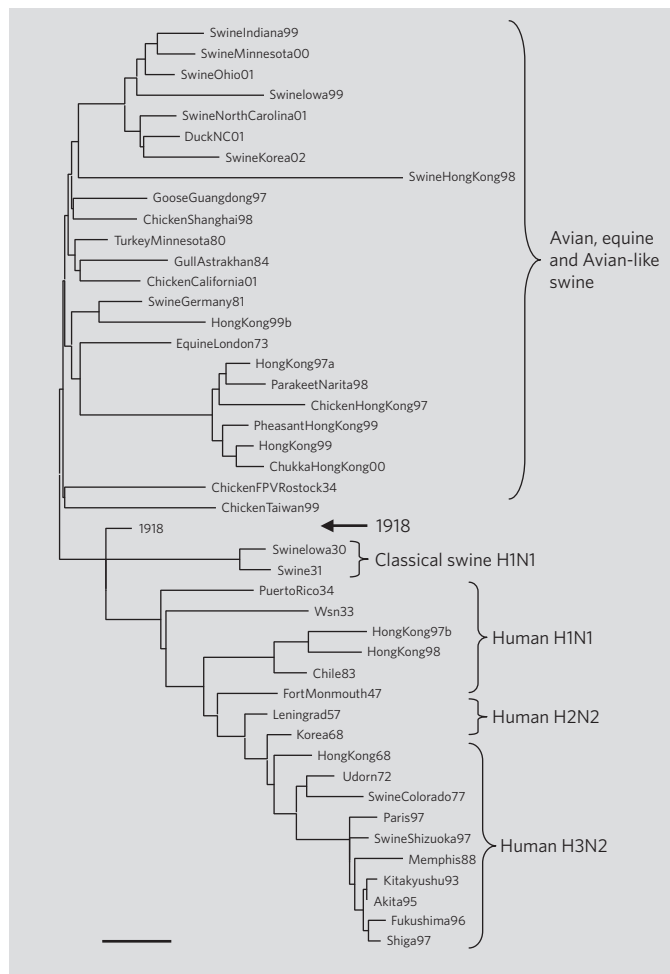


Figure 1 | Neighbour-joining analysis using MEGA 2.1 software¹⁴ of influenza A viral amino-acid sequences of the PB2 polymerase subunit. The arrow designates the 1918 sequence. The scale is proportion of differences.

causing the 1918 pandemic, or it could have infected humans and adapted to them directly several years before the 1918 pandemic. But as no direct precursors of the 1918 virus have been identified, we cannot at present separate these two possibilities. We find the second more plausible, given what we now know about recent human infection with the avian H5N1 influenza virus¹³.

Jeffery K. Taubenberger*, **Ann H. Reid*†**, **Raina M. Lourens*‡**, **Ruixue Wang***, **Guozhong Jin***, **Thomas G. Fanning***

*Armed Forces Institute of Pathology, Department of Molecular Pathology, Rockville, Maryland 20850, USA

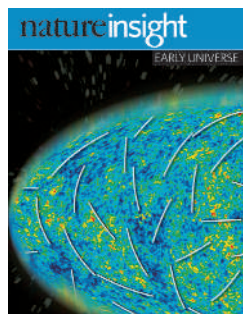
e-mail: taubenberger@afip.osd.mil

†Board on Life Sciences, The National Academies, Washington DC 20001, USA

‡Roy J. and Lucille A. Carver College of Medicine, University of Iowa, Iowa City, Iowa 52242, USA

1. Taubenberger, J. K. *et al.* *Nature* **437**, 889–893 (2005).
2. Reid, A. H., Taubenberger, J. K. & Fanning, T. G. *Nature Rev. Microbiol.* **2**, 909–914 (2004).
3. Taubenberger, J. K., Reid, A. H., Janczewski, T. A. & Fanning, T. G. *Phil. Trans. R. Soc. Lond. B* **356**, 1829–1839 (2001).
4. Gorman, O. *et al.* *J. Virol.* **65**, 3704–3714 (1991).
5. Gibbs, M. J. & Gibbs, A. J. *Nature* **440**, doi:10.1038/nature04823 (2006).
6. Antonovics, J., Hood, M. E. & Baker, C. H. *Nature* **440**, doi:10.1038/nature04824 (2006).
7. Reid, A. H., Fanning, T. G., Hultin, J. V. & Taubenberger, J. K. *Proc. Natl Acad. Sci. USA* **96**, 1651–1656 (1999).
8. Reid, A. H., Fanning, T. G., Janczewski, T. A. & Taubenberger, J. K. *Proc. Natl Acad. Sci. USA* **97**, 6785–6790 (2000).
9. Schultz, U., Fitch, W., Ludwig, S., Mandler, J. & Scholtissek, C. *Virology* **183**, 61–73 (1991).
10. Brown, I. H. *et al.* *J. Gen. Virol.* **78**, 553–562 (1997).
11. Kawaoka, Y., Krauss, S. & Webster, R. G. *J. Virol.* **63**, 4603–4608 (1989).
12. Reid, A. H., Fanning, T. G., Janczewski, T. A., Lourens, R. M. & Taubenberger, J. K. *J. Virol.* **78**, 12462–12470 (2004).
13. Peiris, J. S. *et al.* *Lancet* **363**, 617–619 (2004).
14. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. MEGA2: Molecular Evolutionary Genetics Analysis software, version 2.1 (Arizona State Univ., Tempe, Arizona, 2001).

doi:10.1038/nature04825

**Cover illustration**

Modified from a detailed picture of the infant Universe. (Original image is courtesy of the WMAP/NASA science team.)

Editor, *Nature*

Philip Campbell

Insights Publisher

Sarah Greaves

Insights Editor

Lesley Anson

Production Editors

Maria Hodges
Rebecca Craven

Senior Art Editor

Martin Harrison

Art Editor

Nik Spencer

Layouts

Jane Walker

Production

Jocelyn Hilton

Marketing

Robin Brown

Editorial Assistant

Laura Shaw

The Macmillan Building,

4 Crinan St

London N1 9XW, UK

Tel: +44 (0) 20 7833 4000

e-mail: nature@nature.com



nature publishing group

EARLY UNIVERSE

Cosmology has come a long way in a rather short time. When I was a graduate student 20 years ago, a galaxy with a high redshift was one at $z \approx 1$ (and quasars were known out to $z \approx 2$). Now we know of many distant galaxies at redshifts up to about 6.5 and there are controversial candidates beyond that. We have moved from a model in which mass is dominated by dark matter to one where an even more mysterious dark energy is the main component of the Universe.

In order to highlight recent advances, *Nature* has commissioned a diverse group of people to write about their subjects and how they are developing our understanding of the Universe. This collection seems particularly timely in the light of the recent release of three years of data from WMAP — the Wilkinson Microwave Anisotropy Probe.

Chuck Bennett discusses mysteries that have arisen as the data become ever more tightly constrained by the evolving 'standard model', while Sean Carroll probes the more philosophical question of why our Universe is the way it is. Volker Springel, Carlos Frenk and Simon White find that the growth of structure in the Universe is a powerful way to compare the predictions of the standard model with observations of clusters of galaxies. And Esther Hu and Len Cowie examine recent work at the level of individual galaxies, which concludes that star formation reached its peak when the Universe was about half its present age. Finally, John Cowan and Chris Sneden look at patterns in the abundance of heavy elements in the oldest stars in our Galaxy to infer what kinds of star preceded them.

We hope that you enjoy this collection.

Leslie Sage, Senior Editor

REVIEWS

1126 Cosmology from start to finish

Charles L. Bennett

1132 Is our Universe natural?

Sean M. Carroll

1137 The large-scale structure of the Universe

Volker Springel, Carlos S. Frenk & Simon D. M. White

1145 High-redshift galaxy populations

Esther M. Hu & Lennox L. Cowie

1151 Heavy element synthesis in the oldest stars and the early Universe

John J. Cowan & Christopher Sneden

nature
insight

Cosmology from start to finish

Charles L. Bennett¹

Cosmology is undergoing a revolution. With recent precise measurements of the cosmic microwave background radiation, large galaxy redshift surveys, better measurements of the expansion rate of the Universe and a host of other astrophysical observations, there is now a standard, highly constrained cosmological model. It is not a cosmology that was predicted. Unidentified dark particles dominate the matter content of our Universe, and mysteries surround the processes responsible for the accelerated expansion at its earliest moments (inflation?) and for its recent acceleration (dark energy?). New measurements must address the fundamental questions: what happened at the birth of the Universe, and what is its ultimate fate?

Observational cosmologists cannot actively perturb the Universe to measure how it responds. Instead, they must settle for the passive collection of a small fraction of the photons that chance to come our way. From this perspective, it is truly remarkable that we should have come so far in determining the nature of our vast Universe. Through the extraordinary efforts of observers and theorists over the past century, we now have a well-specified standard model of cosmology that fits a precise and highly constraining set of measurements¹.

The standard model includes the Big Bang theory: 13.7 billion years ago the Universe was hot and dense, and it has expanded and cooled ever since. Observations also specify a precise census of the types of matter and energy in the Universe. The content of the Universe has evolved. It began dominated by an energy field with a negative pressure, which drove an early period of accelerated expansion ('inflation'). It was then dominated by radiation, and later by matter. Recently it has again become dominated by a negative pressure energy that is driving a slower accelerated expansion. Most of the current matter content is in the form of some kind(s) of non-baryonic dark matter. Familiar atomic matter is less than 5% of the mass-energy content of the Universe. Although the standard model of cosmology includes inflation, this idea is not as firmly established as the Big Bang theory. The simplest version of inflation provides a mechanism that simultaneously explains several of the observed properties of the Universe.

There remain three primary cosmological mysteries that must now be confronted. (1) What is the non-baryonic dark matter? (2) Did inflation happen? (3) What is causing the recent apparent acceleration? I discuss each of these below, with a focus on the second and third: the early and late time epochs of accelerated expansion of the Universe.

What is the non-baryonic dark matter?

Non-baryonic dark matter does not emit, absorb or scatter light. There is compelling astrophysical evidence that non-baryonic dark matter is about 5 times as abundant as atoms (baryons). This conclusion relies on the deduced presence of matter, based on its observed gravitational effects, in situations where there is an absence of any interaction with light (such as an immunity of the matter to effects of radiation pressure). For the large-scale structure of the Universe to have evolved in the observed manner (see also the review in this issue by Springel, Frenk and White, p. 1137) the non-baryonic matter must have negligible random motion. We therefore refer to it as cold dark matter (CDM).

Astrophysical observations demand the existence of CDM (or something with similar properties²), but it will be difficult at best for astrophysical observations to identify the specific dark matter particle (or particles). High-energy colliders — such as the future Large Hadron Collider (<http://lhc.web.cern.ch/lhc>) and the proposed International Linear Collider (<http://www.linearcollider.org/cms>) — and ground-based direct detection experiments are much more likely to lead to a definitive identification^{3,4}. Candidates include a $\sim 10^{-5}$ eV axion and a ~ 100 GeV neutralino^{5,6}.

Did inflation happen?

Unlike efforts to identify the non-baryonic dark matter, the measurements required of the two epochs of accelerated expansion of the Universe are likely to come from astrophysical observations.

Although the Big Bang concept of a Universe that adiabatically cools as it expands is extremely well established, observational data currently cannot tightly constrain models of the physics of the very early Universe. Inflation theory is a paradigm in which the early Universe undergoes a brief period of exponential expansion, causing initial quantum perturbations to grow into macroscopic scalar density perturbations and tensor gravitational wave perturbations. The inflationary scheme^{7–9} has been the working model of the very early Universe for over 20 years, and it has had considerable successes.

The simplest version of inflation provides a single mechanism that simultaneously resolves several cosmological questions: why is the Universe so close to flat (euclidean) even though this is an unstable point, requiring extreme fine-tuning? Why do we not observe magnetic monopoles or other topological defects? Why is the Universe so isotropic (in other words, why is the cosmic microwave background temperature very nearly the same in every direction)? What is the source of the perturbations seen as anisotropy in the cosmic microwave background (CMB) radiation in Fig. 1?

If there were ~ 60 e-foldings (that is, a factor of 10^{26}) of inflationary expansion, then the Universe was driven to be nearly flat, removing any need for fine-tuned initial conditions. We would observe a nearly flat Universe today regardless of its initial curvature. The inflationary expansion would have driven the density of magnetic monopoles to become negligible today, explaining their apparent absence. Finally, our entire observable portion of the Universe would have inflated from an initially small causally connected region, thereby ensuring a high degree of isotropy today.

¹Department of Physics and Astronomy, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, USA.

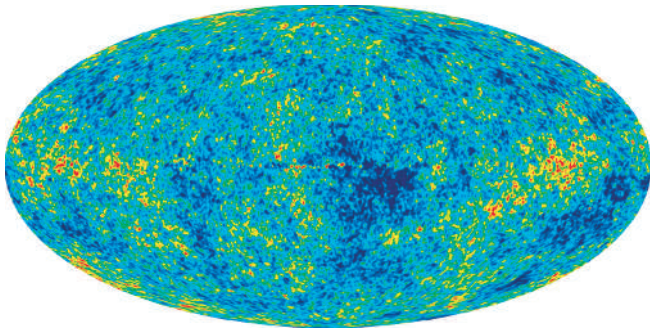


Figure 1 | A map projection of the full-sky CMB anisotropy observed by the Wilkinson Microwave Anisotropy Probe (WMAP)¹¹. The colours represent $\sim 0.001\%$ ($\sim 30 \mu\text{K}$) deviations about the 2.725 K mean temperature of the radiation. Red is warmer and blue is cooler. The largest-scale anisotropy is primordial, whereas smaller-scale anisotropy is from later interactions of the photon–baryon fluid. The coordinates are such that the plane of our Milky Way Galaxy lies along the central horizontal of the map, and the Galactic centre is at the centre of the map. Interfering microwave signals from our own Milky Way have been largely suppressed in this view by use of WMAP's multi-frequency data, which help to distinguish the source of the radiation by its frequency spectrum. (Courtesy of the WMAP Science Team.)

The inflationary scalar and tensor perturbations imprint temperature and polarization anisotropy on the CMB radiation. The simplest form of inflation predicts a nearly scale-invariant power spectrum of scalar primordial gravitational fluctuations, with equal gravitational potential fluctuations on all physical scales. This is described by a power spectrum of scalar (density) fluctuations $P(k) = Ak^{n_s}$, with $n_s = 1$ (the Peebles–Harrison–Zeldovich spectrum). The WMAP sky maps (Fig. 1) show the imprint of these scalar fluctuations on the CMB. The power spectrum of the fluctuations^{10,11}, seen in Fig. 2, displays the peaks and dips that result from the adiabatic gaussian density fluctuations predicted by inflation theory. (For adiabatic fluctuations, the density fluctuations of individual species vary in proportion to the total density variation, whereas for isocurvature fluctuations the densities of individual species vary such that the total density is constant.) Inflation sets the initial conditions for the peaks and dips, which arise from baryon acoustic oscillations in the photon–baryon fluid at the epoch of decoupling (at redshift $z = 1,089$). These oscillations are coherent across the sky because inflation synchronized their temporal development; inflation-generated fluctuations were driven beyond the causal horizon one wavenumber at a time. Later, they entered back into the horizon in a ‘last out, first in’ sequence. As specific modes re-enter the horizon, baryon acoustic oscillations are initiated.

Inflation explains why the Universe is nearly flat, why the CMB temperature is so isotropic and why we see no monopoles. It provides a mechanism by which to generate gaussian adiabatic perturbations with nearly equal gravitational potential perturbations at all angular scales ($n_s \approx 1$).

Measurement of the scalar spectral index, n_s , constrains acceptable inflation models. WMAP data (<http://lambda.gsfc.nasa.gov>), combined with other cosmological data, give $n_s = 0.97 \pm 0.03$ and a Universe within 2% of flat (euclidean), $\Omega_{\text{tot}} = 1.02 \pm 0.02$ (ref. 1).

The physics of accelerated expansion

General relativity¹² and the cosmological principle (the observationally well-supported idea that the Universe is homogeneous and isotropic on large scales) give the acceleration equation

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2} (\rho c^2 + 3p) \quad (1)$$

where $a(t)$ is the dimensionless scale factor as a function of time, $\ddot{a}$ is its second time derivative, G is Newton's constant, ρc^2 is the energy density, and p is pressure. In conditions where negative pressure dominates, an accelerated period of expansion may be expected.

The rate of expansion is given by the Hubble parameter, H , in the Friedmann equation

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3c^2} - \frac{kc^2}{R_c^2 a^2} \quad (2)$$

The last term is a space curvature term with $k = -1, 0$ or $+1$ and R_c is the radius of curvature. Curvature can be largely ignored for inflation because the exponential expansion drives that term towards zero, as discussed above.

With the equation of state $p = w\rho c^2$ the Friedman equation becomes

$$\frac{\ddot{a}}{a} = -\frac{1}{2} H^2 (1+3w) \quad (3)$$

For relativistic matter, $w = 1/3$. For non-relativistic matter, $w = 0$. It is apparent that the expansion of the Universe will accelerate when $w < -1/3$ ($p < -\rho c^2/3$).

In the simplest version of the inflationary paradigm a single scalar field dominates the energy density of the Universe. This scalar field is called the inflaton field, ϕ , and its corresponding effective potential is $V(\phi)$. The pressure of the field is $p(\phi) = \frac{1}{2}\dot{\phi}^2/\hbar c^3 - V(\phi)$ and the total energy is $E(\phi) = \frac{1}{2}\dot{\phi}^2/\hbar c^3 + V(\phi)$. The equation of state parameter is then

$$w = \frac{\frac{1}{2}\dot{\phi}^2/\hbar c^3 - V(\phi)}{\frac{1}{2}\dot{\phi}^2/\hbar c^3 + V(\phi)} \quad (4)$$

To achieve the acceleration condition $w < -1/3$ implies that $\dot{\phi}^2 < \hbar c^3 V(\phi)$. To obtain the degree of inflation required for the observed properties of our Universe (nearly isotropic and flat) the field must evolve (‘roll’) slowly such that the potential dominates over the kinetic terms for a sufficient number of e-foldings of expansion. Acceptable slow-roll single-field inflation requires $\dot{\phi}^2 \ll \hbar c^3 V(\phi)$.

Inflation provides a single physical mechanism that explains and has even predicted observational results. Yet we have not demonstrated that inflation actually happened, nor do we know the identity of the inflaton field or the functional form of its effective potential. It is imperative that we seek answers to the following questions. (1) Did inflation happen? (2) If so, what is ϕ ? (3) What is the functional form of $V(\phi)$?

The obvious candidate for the inflationary scalar field is a Higgs-like field at the Grand Unified Theory (GUT) energy scale. Although this provides a speculation for the identity of ϕ , it does not reveal the form of $V(\phi)$. It does, however, provide a testable prediction.

Inflation generically produces gravitational waves (tensor fluctuations). Gravitational waves leave the horizon during inflation with the same dimensionless strain amplitude at each wavelength. Suggested forms of the effective field energy density $V(\phi)$ (in the common particle physics units of GeV^4) include exponentials and power laws. If the amplitude of the energy scale of inflation is M (in GeV), then we can write $M \propto V^{1/4}$. Measurements of the polarized CMB power spectrum determine Q_T^2 , the tensor power referred to the quadrupole angular scale. This is directly proportional to the energy scale of inflation¹³, $M_*/m_{\text{Pl}} = 1.15 \langle Q_T^2 \rangle^{1/4}$ where m_{Pl} is the Planck mass and M_* represents inflationary potential values when the scale that fixed the CMB quadrupole crossed outside the horizon. This expression conveys the central imperative of a measurement of the CMB tensor power, Q_T^2 . The scalar CMB perturbation power (referred to the quadrupole scale) has been measured by COBE¹⁷ to be $Q_S^2 \approx (17 \mu\text{K}/2.725 \text{ K})^2 \approx 4 \times 10^{-11}$. The tensor-to-scalar ratio, $r = \langle Q_T^2 \rangle / \langle Q_S^2 \rangle$, has already been constrained by measurement. WMAP reported upper limits of $r < 0.9$ (95% CL at the 0.002 Mpc^{-1} scale, with a somewhat different definition of r)¹. Because the scalar power has been measured, and the tensor-to-scalar ratio has a measured upper limit, the tensor power is also constrained. The connection can be expressed as $M_*/m_{\text{Pl}} \approx 3 \times 10^{-3} r^{1/4}$. The amplitude of the inflaton potential can be expressed in terms of the tensor-to-scalar ratio, and specific inflationary models have different tensor-to-scalar ratios, r .

Compelling (least fine-tuned) inflation models¹⁵ have $M_* > 1 \times 10^{16} \text{ GeV}$ in a limited region of the (n_s, r) plane with $r > 10^{-2}$. Future CMB temperature measurements will significantly improve

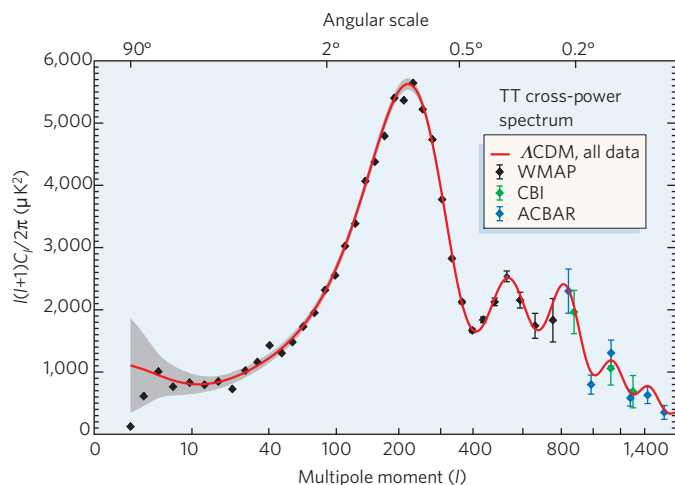


Figure 2 | The angular power spectrum of the WMAP data, supplemented by CBI and ACBAR data at high multipole moments (smaller angular scales)^{10,11}. The red model curve fits these and other cosmological data. The observed series of peaks and dips were successfully predicted by adiabatic inflation theory, which seeds the baryon acoustic oscillations. C_l , angular power spectrum; TT, temperature–temperature cross-power spectrum. (Courtesy of the WMAP Science Team.)

on the current scalar measurements of n_s , and future CMB polarization measurements will test inflation by either detecting or placing upper limits on Q_T^2 , r and M_* . Ongoing CMB tests for deviations from gaussianity are also an important check on inflationary models.

The compelling inflation model of a single scalar field at the GUT energy scale can be put to a rigorous observational test. The CMB community recently concluded that this measurement is practical to carry out within the next decade from a space platform¹⁶. A careful space-borne measurement will not be limited by detector sensitivity but by systematic measurement errors and by the polarized astrophysical foregrounds. The design challenge and goal is to suppress systematic measurement errors to below the foreground limits. Multifrequency measurements are needed to discriminate the foreground polarization from the tensor perturbations. The degree to which this can be done establishes the foreground limit. If single-field GUT-scale inflation happened, the space mission will detect its unique polarized signature, and achieve a milestone in science. If the polarized signature is not detected, our simplest, 25-year-old view of the early Universe would be ruled out. More exotic models (perhaps with extra dimensions, branes¹⁷, string theory, and complex versions of inflation) would be developed and compete in terms of ‘naturalness’ and simplicity. We are at the threshold of a definitive test of simple GUT-scale inflation, our most compelling model of the early Universe.

Why the recent accelerated expansion?

Equation (1) shows that the acceleration of the Universe is to be expected during epochs when the pressure is sufficiently negative, as with inflation. There is good evidence that the criterion is again met in the recent Universe. What is causing the negative pressure in the Universe today? Why has this new epoch of accelerated expansion only recently begun?

The first evidence of the recent accelerated expansion came from observations of exploding white dwarf stars. A white dwarf star is the remnant of a small to moderate star. Its core, predominantly carbon and oxygen, is supported by degenerate electron pressure. A lone white dwarf fades by radiating its heat. However, if a white dwarf has a companion star, mass can transfer from the companion to the white dwarf until the degenerate electron pressure can no longer support the white dwarf. This occurs at the Chandrasekhar limit, at a mass of 1.4 times that of our Sun. The resulting implosion detonates the carbon core, which causes the star to explode as a type Ia supernova (SN Ia). Although the details of the

explosion process are not known, the common nature of this physical process suggests that all SNe Ia explosions may be similar. As a result, in 1979 Sterling Colgate suggested¹⁸ that SNe Ia explosions of white dwarf stars be used as standard candles for determining distances.

The expansion of the Universe should decelerate if it is dominated by baryonic and cold dark matter. Two independent groups observed SNe Ia in an attempt to measure the expected slowing of the expansion of the Universe. Instead, in 1998, both groups made the surprising discovery that the expansion has recently been accelerating^{19,20}. Before this discovery, the issue surrounding stellar ages already hinted at the need for dark energy²¹.

It is not surprising that the first detection of dark energy was made with supernovae as there was high value to the measurement of each individual object. Even now, the statistical sample is relatively small. Other methods have much less value per object, relying more on overall statistics. It was not until the 1998 supernovae results were corroborated by independent CMB and large-scale structure measurements in 2003 that the accelerated expansion became undeniable. In 2003, *Science* declared this as the breakthrough science of the year²²: “...many cosmologists hesitated to embrace the bizarre idea of dark energy... Linger doubts about the existence of dark energy and the composition of the Universe dissolved when the WMAP satellite took the most detailed picture ever of the cosmic microwave background (CMB)... physicists superimposed the galaxy-clustering data of SDSS on the microwave data of WMAP and proved — beyond a reasonable doubt — that dark energy must exist.” Meanwhile, the supernovae samples have steadily grown in numbers and redshift²³.

The reason for the recent accelerated expansion of the Universe is still a mystery, but it suggests an unexpected new negative pressure contributor to the mass-energy of the Universe in equation (1). But equation (1) is derived from general relativity so an alternative explanation is that general relativity is flawed.

One modification to the field equation was made by Einstein himself: the introduction of a cosmological constant. Einstein introduced this as a gravitational repulsive term to force the Universe to be static. When Hubble later demonstrated that the Universe expands, the idea was dropped. The cosmological constant may play the antigravity role that drives the current accelerated expansion. The effective equation of state parameter of the cosmological constant is $w = -1$. The zero-point energy of the vacuum could act as a cosmological constant. Another possibility is that there is a universal evolving scalar field driving the recent accelerated expansion, with equation of state parameter of $w < -1/3$.

All three categories of explanation for the recent accelerated expansion of the Universe have been extensively studied: modifications to general relativity^{24–29}, a cosmological constant^{30–32} and a universal evolving scalar field^{31–36}. All have problems. It has turned out to be remarkably difficult to modify general relativity on the largest scales in ways that are not already in conflict with other measured behaviour^{37,38}. Investigators are pursuing several active lines of research into gravity in the context of extra dimensions³⁹. Perhaps gravity leaks into extra dimensions. In seeking a unified theory of all of the forces of nature, this might help us to understand gravity, which has stubbornly resisted unification.

As we have seen, an evolving scalar field is the standard explanation for inflation, so it might seem reasonable for the recent acceleration as well, but that is not the case. During inflation, the characteristic energy is naturally associated with an existing symmetry at the GUT scale. The GUT scale is the energy where the electromagnetic, weak, and strong forces become indistinguishable at $\sim 10^{16}$ GeV (or an energy density of 10^{64} GeV⁴ in the usual particle physics units). A proposed ‘dark energy’ scalar field at the extraordinarily low energy of 10^{-3} eV (10^{-48} GeV⁴) to explain the recent accelerated expansion presents a tremendous hierarchy problem. In general, scalar fields tend to be heavy (high energy). When renormalized, scalar fields tend to become even heavier. This is acceptable for inflation because the Universe was in a very high energy state at that epoch, but it seems highly unnatural for the recent Universe.

The energy density of the Universe has been in decline since the end of inflation. It might seem natural that it has declined to the point where the

Box 1 | Type Ia supernovae

Dark energy was initially discovered by two independent groups measuring the expansion rate of the Universe with type Ia supernovae⁵⁰ as distance indicators^{19,20}. The supernovae are used to measure the luminosity distance, defined by $d_L = \sqrt{L/(4\pi F)}$, where L is the intrinsic luminosity of a supernova and F is the observed flux, both within a passband. This can be rewritten in favour of the apparent magnitude, m (as a logarithmic measure of the flux), and the absolute magnitude, M (as a logarithmic measure of the luminosity). We then have the (extinction-corrected) distance modulus $\mu \equiv m - M = 5 \log_{10}(d_L/1 \text{ Mpc})$. Figure 3 shows the type Ia supernova measurements as a function of redshift, where Hubble Space Telescope (HST) data dominate for $z > 1$.

Although SNe Ia are often referred to as 'standard candles', it would be more correct to call them 'standardizable candles' as they are known to show variations in luminosity. These variations can be regressed out to a great extent^{51–53}. Apparently, there are significant differences between type Ia SNe explosions that affect their yields of radioactive nickel. These differences are believed to be responsible for peak output variations of a full range factor of 6, with a dispersion of ~40%. The use of empirical diagnostics (as well as dropping ~5% of pathologically non-standard SNe Ia) can reduce the remaining dispersion to ~15% in the current observed samples. The dispersion for a much larger survey at higher redshifts is unknown.

To realize improvements in SNe-derived cosmology from increased sample size, the systematic errors must become proportionately smaller. Yet, as the SNe sample sizes grow and the observed distribution extends to greater look-back times, adverse systematic effects due to evolution, dust scattering, and gravitational lensing will be increasingly important. An oft-stated desire is to reach an empirical understanding of the SNe Ia luminosity to the 1% level, including calibration uncertainties. This level of accuracy may become possible, but will not be easy to achieve.

small energy density of the vacuum is now dominant for the first time. Why now? The timing seems an uncomfortable coincidence. More troubling, however, is the energy scale. This fundamental problem existed before the discovery of the recent accelerated expansion. It had been assumed that there is an exact cancellation in an as yet undetermined theory that sets the vacuum energy exactly to zero. To associate the dark energy with the vacuum energy is to demand a huge, but not perfect, chance cancellation to a part in 10^{124} of the Planck energy density (at 10^{76} GeV^4) to the vacuum energy density (10^{-48} GeV^4). This seems an unreasonable fine-tuning.

Another possibility is that the cosmological constant is not set by theory at all, but is an environmental variable that varies from one region of the Universe (or multiverse) to the next. An anthropic argument can be made that we necessarily live within the small subset of universes that are capable of producing galaxies, planets and observers, so a small but non-zero cosmological constant becomes likely⁴⁰. For example, a much larger cosmological constant would drive an accelerated expansion that prevented bound objects from forming. Our existence creates a prior observational bias.

Observational tests of dark energy

As there is no compelling explanation for the recent accelerated expansion, observational tests might have profound implications for physics. Despite the fact that dark energy recently began to dominate the Universe, it is remarkably difficult to measure its properties precisely.

The determination of the nature of dark energy will be difficult and will inevitably rely on the combined information from different independent measurements. The lack of uniform standards is already a problem. Some estimates constrain dark energy while assuming an exactly flat geometry of the Universe, and others marginalize over measured geometric limits. Also, dark energy limits are given in terms of different parameter definitions. For example, fitting a set of data to $w(z) = w_0 + w_1 z$ will produce a different result from a fit to $w(z) = w_0 + w_1 \ln(1 + z)$, or $w(z) = w_0 + w_1 z/(1 + z)$, or $w(z) = w_0 + w_1 \sin(\pi z/(1 + z))$.

Current observations provide the tightest uncertainty, σ_w , at $w(z \approx 0.2)$.

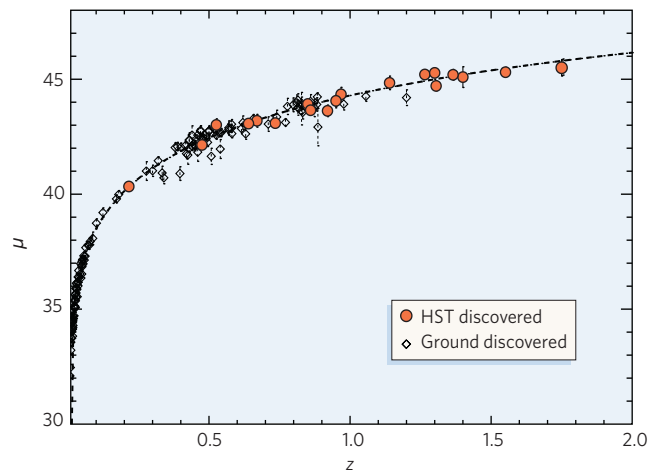


Figure 3 | The supernova Ia Hubble diagram⁴⁸. The extinction-corrected distance modulus is shown as a function of redshift. The ground-discovered supernovae Ia are shown as diamonds. HST-discovered SNe Ia are shown as filled symbols. The model curve is the best fit cosmology of $\Omega_m = 0.29$ and $\Omega_L = 0.71$, assuming a flat Universe and a constant $w = -1$ equation of state parameter. Supernovae Ia observations suggest that a transition from recent acceleration to past deceleration occurred at $z = 0.46 \pm 0.13$.

These constraints will improve with data from current and planned future efforts. Space-based efforts will not significantly improve on the ground-based limits on $\sigma_w(z \approx 0.2)$, but can place better limits on the time evolution of $w(z)$. For example, they could improve $\sigma_w(z > 0.8)$ by substantially larger sampling of the higher-redshift Universe. Roughly speaking, ground-based efforts determine the recent value of w_0 , while space-based efforts provide leverage on its variability. The key goal of the planned Joint Dark Energy Mission (JDEM) is to improve significantly on dark energy limits at the higher redshifts.

Results of ongoing cosmological measurements with no direct connection to dark energy will indirectly assist in determining its nature by constraining cosmological parameters that are partially degenerate with dark energy parameters. Examples include measurements of the geometry of the Universe and the matter density. Future CMB measurements, such as the Planck Surveyor space mission and the ground-based South Pole Telescope and Atacama Cosmology Telescope should help to place tighter limits on these and other cosmological parameters.

Some of the main observational approaches to dark energy include type Ia SNe (see Box 1 and Fig. 3), baryon acoustic oscillations (see Box 2 and Fig. 4), weak gravitational lensing and cluster studies.

Weak gravitational lensing

The weak gravitational lensing of distant galaxies by the intervening matter distribution is a promising method for determining the three-dimensional distribution of matter in the Universe⁴¹. From this, the expansion rate and the growth rate of structure with redshift can be deduced. Because these observations rely on the distortion of the apparent shapes of galaxies, high resolution and a stable point spread function are needed. To obtain a good statistical sample, a wide sky area is also needed, with accurately measured redshifts. Cosmological results so far⁴² have been limited by the observational challenge. The first generation of observations, using existing instruments, was limited by some combination of instrument, telescope and atmospheric motions. New instrumentation, specifically designed to minimize weak lensing systematic errors, will bring advances. The US National Academy of Sciences recommended the construction of a Large Synoptic Survey Telescope to carry out these (and other) observations, and efforts are under way⁴³ (see also http://www.lsst.org/Science/fs_darkenergy.shtml).

Galaxy clusters

Cosmological observations called Alcock–Paczynski tests⁴⁴ can be used to probe the dark energy. The idea is to identify a class of objects that

Box 2 Baryon acoustic oscillations

Cosmological perturbations in the early Universe excite sound waves in the photon-baryon fluid⁴⁹. These baryon acoustic oscillations, seen in the CMB in Fig. 2, define a standard ruler whose length is the distance that sound can travel, at a speed of $\sim c/\sqrt{3}$, before decoupling^{55,56}. The 'sound horizon' scale is also imprinted on the distribution of matter in the Universe. Recently, these oscillations have been seen in the spatial distribution of galaxies^{57,58}, as shown in Fig. 4. Baryon acoustic oscillations are well understood, involving classical physics on a scale where linear perturbations are dominant. Nature has provided a standard measuring rod.

Measurements of the apparent angular (transverse) size of oscillations on the sky probe the relationship between angular diameter distance and redshift. The angular diameter distance is $d_A \equiv l/\theta = d_l/(1+z)^2$ where l is the actual size, and θ is the apparent angular size. The angular diameter distance depends on the co-moving distance and the space curvature.

In a large three-dimensional galaxy redshift survey the oscillation scale can be measured radially (Δz) and in transverse angle, thereby providing the instantaneous expansion rate, $H(z)$ as well as an integral over $H(z)$. The geometry and expansion history of the Universe can be derived from the tangential and radial components, and dark energy constrained^{59–63}.

Powerful new constraints on the nature of the dark energy can be realized if the apparent acoustic scale in the statistical distribution of galaxies is measured over large effective volume of the Universe. Potential systematic errors include the effects of nonlinearity, galaxy bias and redshift space distortions. These are smooth functions over the scales of interest, which exceed 30 Mpc, so they are not great impediments to fitting oscillatory features in the power spectrum. Existing simulations show residual errors to be small^{64–70}, and they can probably be further reduced with more observational and simulation experience.

The main challenge for the baryon acoustic oscillation method lies in obtaining the necessary large effective volume surveys. High-redshift surveys from the ground are limited by bright airglow, but advances are possible. Options for future ground-based and space-based observations have been discussed^{71–73}. We do know that there is a sufficient population of high-redshift galaxies to sample⁷⁴.

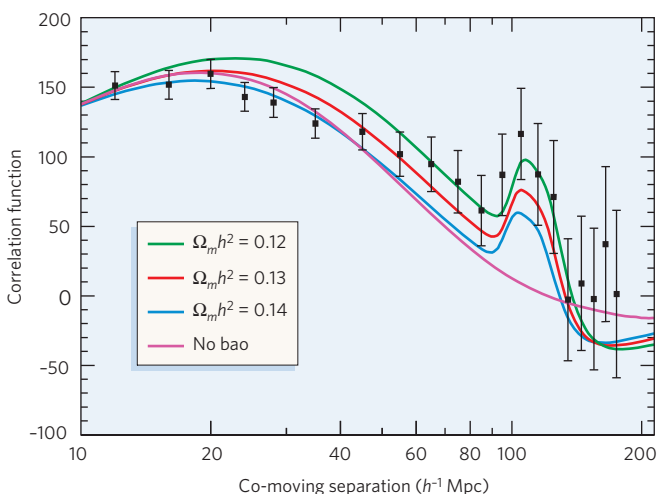


Figure 4 | The galaxy redshift-space correlation function reveals baryon acoustic oscillations. The data are from red luminous galaxies in the 3816 square degree Sloan Digital Sky Survey (SDSS)⁴⁹. The baryon oscillation signal visible at $\sim 100 h^{-1}$ Mpc (where h is the Hubble constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$) is detected at 3.4σ . The green, red and blue models have $\Omega_m h^2 = 0.12$, 0.13 and 0.14 , respectively. The magenta curve is a CDM model with no baryon oscillations. The acoustic peak gives the ratio of the distances to $z = 0.35$ and $z = 1,100$ to 4% fractional accuracy. The absolute distance to $z = 0.35$ is determined to 5% accuracy. The full correlation function shape gives $\Omega_m h^2$ to 8% and agrees with the value from the CMB. The co-moving sound horizon scale is $\sim 150 h^{-1}$ Mpc.

are isotropic, where radial (redshift) properties are equivalent to their tangential (angular) properties, at least statistically. Another example uses galaxy clusters, the largest bound structures in the Universe. Large statistical samples of clusters will be accumulated, including optical, X-ray, and CMB Sunyaev–Zeldovich measurements. Cluster counts and cluster evolution as a function of redshift are sensitive to changes in dark energy⁴⁵.

The path forward

The Universe experienced an accelerated expansion at both early and late times. The early inflation could not have been due to a cosmological constant because it stopped, but the recent inflation may continue forever. Both inflationary epochs have been explained in terms of scalar fields (at vastly different energy scales), but it is sobering to recognize that there is no compelling evidence yet for the existence of any fundamental scalar fields. Perhaps the two periods of inflation have some intimate connection that we have not yet recognized.

Although one can construct arbitrarily complex and multiple potentials to make inflationary models predict an almost limitless range of behaviour, only the simplest form is compelling because only it simultaneously explains a wide range of observed phenomena with a minimum of free parameters and little fine-tuning. The simplest inflation models¹⁵ give $n_s < 0.98$ and $r > 10^{-2}$ and generate an isotropic, very nearly flat Universe, with a gaussian adiabatic spectrum of perturbations.

It is an observational imperative to determine r , n_s and flatness with greater precision. Efforts aimed at more precise measurements of the scalar power spectrum index are under way from space (WMAP, Planck) and from the ground (the Atacama Cosmology Telescope, the South Pole Telescope, and others).

CMB measurements have already led to constraints on the shape of the inflaton potential, and future observations can powerfully extend these constraints. This has been strongly emphasized⁴⁶: "Discovery of inflation-produced gravity waves is a smoking gun signature of inflation, immediately identifies the epoch and scale of inflation and reveals information about the inflationary potential."

It is imperative to search for the gravitational waves from the early Universe by means of the polarized CMB signal, and to measure r from that detection. An eagerly anticipated space mission, commonly known as CMBPol, was recommended by the US National Academy of Sciences⁴⁷ to "[m]easure the polarization of the cosmic microwave background with the goal of detecting the signature of inflation." NASA has the mission on its roadmap and has funded mission concept studies. A CMB task force has reported a community consensus that a space mission is required.

We have no compelling model for the recent epoch of accelerated expansion. Ground-based dark energy measurements give the equation of state at $z \approx 0.2$. Space-borne measurements from JDEM will focus on higher redshifts and will limit the possible evolution of $w(z)$. There can never be an exact verification of the cosmological constant because measurements will always have uncertainties that allow for a slowly evolving scalar field with $w_0 \approx -1$. One might have hoped that these scalar field models could be dismissed as excessively fine-tuned, but this has not been established and may not be true. Of course, a clear deviation from cosmological constant behaviour would be a most important discovery.

The standard cosmological model of our Universe allows astrophysical objects to grow. This is not the case in most cosmological models. In our Universe, space-time is an enormous and mostly lifeless vacuum. Stars, planets and life are transitory, squeezed between the two epochs of accelerated expansion. To understand the habitability of our Universe (and others) we must answer our most fundamental cosmological questions. How did the Universe begin? How will it end? Determining answers to these questions is the goal of current intensive research and is the objective of the JDEM and CMBPol space missions. Remarkably, the necessary measurements are now within our capabilities.

Note added in proof: New WMAP results were issued when this article was in the press (see <http://lambda.gsfc.nasa.gov>).

1. Spergel, D. N. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters. *Astrophys. J. Suppl. Ser.* **148**, 175–194 (2003).
2. Ostriker, J. P. & Steinhardt, P. J. New light on dark matter. *Science* **300**, 1909–1913 (2003).
3. Gaitskell, R. J. Direct detection of dark matter. *Annu. Rev. Nucl. Part. Sci.* **54**, 315–359 (2004).
4. Bertone, G., Hooper, D. & Silk, J. Particle dark matter: evidence, candidates and constraints. *Phys. Rep.* **405**, 279–390 (2005).
5. Ellis, J., Ferstl, A. & Olive, K. A. Theoretical aspects of dark matter detection. *Space Sci. Rev.* **100**, 235–246 (2002).
6. Ellis, J., Olive, K. A., Santos, Y. & Spanos, V. C. Update on the direct detection of supersymmetric dark matter. *Phys. Rev. D* **71**, 095007 (2005).
7. Guth, A. H. Inflationary universe: A possible solution to the horizon and flatness problems. *Phys. Rev. D* **23**, 347–356 (1981).
8. Albrecht, A. & Steinhardt, P. J. Cosmology for grand unified theories with radiatively induced symmetry breaking. *Phys. Rev. Lett.* **48**, 1220–1223 (1982).
9. Linde, A. D. Coleman–Weinberg theory and the new inflationary universe scenario. *Phys. Lett. B* **114**, 431–435 (1982).
10. Hinshaw, G. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: The angular power spectrum. *Astrophys. J. Suppl. Ser.* **148**, 135–159 (2003).
11. Bennett, C. L. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Preliminary maps and basic results. *Astrophys. J. Suppl. Ser.* **148**, 1–27 (2003).
12. Einstein, A. Die Grundlage der allgemeinen Relativitätstheorie. *Ann. Phys.* **49**, 769–822 (1916).
13. Turner, M. S. & White, M. Dependence of inflationary reconstruction upon cosmological parameters. *Phys. Rev. D* **53**, 6822–6828 (1996).
14. Bennett, C. L. *et al.* Four-year COBE DMR cosmic microwave background observations: Maps and basic results. *Astrophys. J. Lett.* **464**, L1–L4 (1996).
15. Boyle, L. A., Steinhardt, P. J. & Turok, N. Inflationary predictions reconsidered. <astro-ph/0507455> (2005).
16. Report of the Task Force on Cosmic Microwave Background Research. <http://www.nsf.gov/mps/ast/tfcr.jsp> (2005).
17. Steinhardt, P. J. & Turok, N. The cyclic model simplified. *New Astron. Rev.* **49**, 43–57 (2005).
18. Colgate, S. A. Supernovae as a standard candle for cosmology. *Astrophys. J.* **232**, 404–408 (1979).
19. Riess, A. G. *et al.* Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116**, 1009–1038 (1998).
20. Perlmutter, S. *et al.* Measurements of Omega and Lambda from 42 high-redshift supernovae. *Astrophys. J.* **517**, 565–586 (1999).
21. Dunlop, J. *et al.* A 3.5-Gyr-old galaxy at redshift 1.55. *Nature* **381**, 581–584 (1996).
22. Seife, C. Illuminating the dark universe. *Science* **302**, 2038–2039 (2003).
23. Astier, P. *et al.* (The SNLS Collaboration) The Supernova Legacy Survey: measurement of Ω_M , Ω_Λ and w from the first year data set. *Astron. Astrophys.* <astro-ph/0510447> (2005).
24. Dvali, G. R., Gabadadze, G. & Porrati, M. 4D gravity on a brane in 5D Minkowski space. *Phys. Lett. B* **485**, 208–214 (2000).
25. Deffayet, C., Dvali, G. R. & Gabadadze, G. Accelerated universe from gravity leaking to extra dimensions. *Phys. Rev. D* **65**, 044023 (2002).
26. Carroll, S. M., Duvvuri, V., Trodden, M. & Turner, M. S. Is cosmic speed-up due to new gravitational physics? *Phys. Rev. D* **70**, 043528 (2004).
27. Nojiri, S. & Odintsov, S. D. Modified gravity with negative and positive powers of curvature: Unification of inflation and cosmic acceleration. *Phys. Rev. D* **68**, 123512 (2003).
28. Arkani-Hamed, N., Cheng, H. C., Luty, M. A. & Mukohyama, S. Ghost condensation and a consistent IR modification of gravity. *J. High Energy Phys.* **5**, 74 (2004).
29. Arkani-Hamed, N., Creminelli, P., Mukohyama, S. & Zaldarriaga, M. Ghost inflation. *J. Cosmol. Astropart. Phys.* **4**, 001 (2004).
30. Weinberg, S. The cosmological constant problem. *Rev. Mod. Phys.* **61**, 1 (1989).
31. Peebles, P. K. E. & Ratra, B. The cosmological constant and dark energy. *Rev. Mod. Phys.* **75**, 559–606 (2003).
32. Zlatev, I., Wang, L. M. & Steinhardt, P. J. Quintessence, cosmic coincidence, and the cosmological constant. *Phys. Rev. Lett.* **82**, 896–899 (1999).
33. Wetterich, C. Cosmology and the fate of dilatation symmetry. *Nucl. Phys. B* **302**, 668–696 (1988).
34. Peebles, P. J. E. & Ratra, B. Cosmology with a time-variable cosmological ‘constant’. *Astrophys. J. Lett.* **325**, L17–L20 (1988).
35. Carroll, S. M. Quintessence and the rest of the world: suppressing long-range interactions. *Phys. Rev. Lett.* **81**, 3067–3070 (1998).
36. Armendariz-Picon, C., Mukhanov, V. & Steinhardt, P. J. Essentials of k-essence. *Phys. Rev. D* **63**, 103510 (2001).
37. Will, C. M. The confrontation between general relativity and experiment. *Pramana* **63**, 731–740 (2004).
38. Bennett, C. L. Astrophysical observations: lensing and eclipsing Einstein’s theories. *Science* **307**, 879–884 (2005).
39. Randall, L. & Sundrum, R. An alternative to compactification. *Phys. Rev. Lett.* **83**, 4690–4693 (1999).
40. Weinberg, S. Anthropic bound on the cosmological constant. *Phys. Rev. Lett.* **59**, 2607–2610 (1987).
41. Refregier, A. Weak gravitational lensing by large-scale structure. *Annu. Rev. Astron. Astrophys.* **41**, 645–668 (2003).
42. Van Waerbeke, L., Mellier, Y. & Hoekstra, H. Dealing with systematics in cosmic shear studies: New results from the VIRMOS–Descart survey. *Astron. Astrophys.* **429**, 75–84 (2005).
43. Hodapp, K. W. *et al.* Design of the Pan-STARRS telescopes. *Astron. Nachr.* **325**, 636–642 (2004).
44. Alcock, C. & Paczynski, B. An evolution free test for non-zero cosmological constant. *Nature* **281**, 358–359 (1979).
45. Haiman, Z., Mohr, J. & Holder, G. Constraints on cosmological parameters from future galaxy cluster surveys. *Astrophys. J.* **553**, 545–561 (2001).
46. Turner, M. S. The new cosmology: mid-term report card for inflation. *Ann. Henri Poincaré* **4**, 5333–5346 (2003).
47. National Research Council. *Connecting Quarks with the Cosmos. Eleven: Science Questions for the New Century* (The National Academies Press, Washington DC, 2003).
48. Riess, A. G. *et al.* Type Ia supernova discoveries at $z > 1$ from the Hubble Space Telescope: Evidence for past deceleration and constraints on dark energy evolution. *Astrophys. J.* **607**, 665–687 (2004).
49. Hu, W., Sugiyama, N. & Silk, J. The physics of microwave background anisotropies. *Nature* **386**, 37–43 (1997).
50. Leibundgut, B. Cosmological implications from observations of Type Ia supernovae. *Annu. Rev. Astron. Astrophys.* **39**, 67–98 (2001).
51. Phillips, M. M. The absolute magnitudes of Type Ia supernovae. *Astrophys. J. Lett.* **413**, L105–L108 (1993).
52. Phillips, M. M. *et al.* The reddening-free decline rate versus luminosity relationship for Type Ia supernovae. *Astrophys. J.* **118**, 1766–1776 (1999).
53. Riess, A. G., Press, W. H. & Kirshner, R. P. A precise distance indicator: Type Ia supernova multicolor light-curve shapes. *Astrophys. J.* **473**, 88–109 (1996).
55. Peebles, P. J. E. & Yu, J. T. Primeval adiabatic perturbation in an expanding universe. *Astrophys. J.* **162**, 815–836 (1970).
56. Eisenstein, D. & Hu, W. Baryonic features in the matter transfer function. *Astrophys. J.* **496**, 605–614 (1998).
57. Eisenstein, D. *et al.* Detection of the baryon acoustic peak in the large-scale correlation function of SDSS luminous red galaxies. *Astrophys. J.* **633**, 560–574 (2005).
58. Cole, S. *et al.* The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications. *Mon. Not. R. Astron. Soc.* **362**, 505–534 (2005).
59. Blake, C. & Glazebrook, K. Probing dark energy using baryonic oscillations in the galaxy power spectrum as a cosmological ruler. *Astrophys. J.* **594**, 665–673 (2003).
60. Seo, H.-J. & Eisenstein, D. J. Baryonic acoustic oscillations in simulated galaxy redshift surveys. *Astrophys. J.* **633**, 575–588 (2005).
61. Eisenstein, D. J., Hu, W. & Tegmark, M. Cosmic complementarity: H_0 and Ω_M from combining cosmic microwave background experiments and redshift surveys. *Astrophys. J. Lett.* **504**, L57–L60 (1998).
62. Linder, E. V. Baryon oscillations as a cosmological probe. *Phys. Rev. D* **68**, 083504 (2003).
63. Hu, W. & Haiman, Z. Redshifting rings of power. *Phys. Rev. D* **68**, 3004 (2003).
64. Angulo, R. *et al.* Constraints on the dark energy equation of state from the imprint of baryons on the power spectrum of clusters. *Mon. Not. R. Astron. Soc. Lett.* **362**, L25–L29 (2005).
65. Coles, P. Galaxy formation with a local bias. *Mon. Not. R. Astron. Soc.* **262**, 1065–1075 (1993).
66. Scherrer, R. J. & Weinberg, D. H. Constraints on the effects of locally biased galaxy formation. *Astrophys. J.* **504**, 607–611 (1998).
67. Springel, V. *et al.* Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature* **435**, 629–636 (2005).
68. White, M. Baryon oscillations. *Astropart. Phys.* **24**, 334–344 (2005).
69. Eisenstein, D. J. & White, M. Theoretical uncertainty in baryon oscillations. *Phys. Rev. D* **70**, 103523 (2004).
70. Glazebrook, K. & Blake, C. A. Measuring the cosmic evolution of dark energy with baryonic oscillations in the galaxy power spectrum. *Astrophys. J.* **631**, 1–20 (2005).
71. Glazebrook, K., Baldry, I., Moos, W., Kruk, J. & McCandliss, S. Monster redshift surveys through dispersive slitless imaging: The Baryon Oscillation Probe. *New Astron. Rev.* **49**, 374–378 (2005).
72. Bassett, B. A., Nichol, R. C. & Eisenstein, D. J. WFMOS — Sounding the dark cosmos. *Astron. Geophys.* **46**, 5.26–5.29 (2005).
73. Hill, G. J. *et al.* in *Ground-Based Instrumentation for Astronomy* (eds Moorwood, A. F. M. & Masanori, I.) *Proc. SPIE* **5492**, 251–261 (2004).
74. Glazebrook, K. *et al.* A high abundance of massive galaxies 3–6 billion years after the Big Bang. *Nature* **430**, 181–184 (2004).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to the author (cbennett@jhu.edu).

Is our Universe natural?

Sean M. Carroll¹

It goes without saying that we are stuck with the Universe we have. Nevertheless, we would like to go beyond simply describing our observed Universe, and try to understand why it is that way rather than some other way. When considering both the state in which we find our current Universe, and the laws of physics it obeys, we discover features that seem remarkably unnatural to us. Physicists and cosmologists have been exploring increasingly ambitious ideas in an attempt to explain how surprising aspects of our Universe can arise from simple dynamical principles.

What makes a situation 'natural'? Ever since Newton, we have divided the description of physical systems into two parts: the configuration of the system, characterizing its particular state at some specific time, and the dynamical laws governing its evolution. For either part of this description, we have an intuitive notion that certain possibilities are more robust than others. When we come across a situation that seems unnatural or finely tuned, physicists seize upon it as a clue pointing towards some underlying mechanism that made it that way. Such clues can occasionally be misleading, but they often serve to guide our thinking about how we can extend our understanding into unknown domains.

For configurations, the concept of entropy quantifies how likely a situation is. If we find a collection of gas molecules in a high-entropy state distributed uniformly in a box, we are not surprised, whereas if we find the molecules huddled in a low-entropy configuration in one corner of the box, we imagine there must be some explanation. For dynamical laws, the concept of naturalness can be harder to quantify. As a rule of thumb, we expect dimensionless parameters in a theory to be of order unity, not too large or too small. Indeed, in the context of effective quantum field theories, the renormalization group gives us some justification for this notion of 'naturalness'¹.

In field theory, the dynamics of the low-energy degrees of freedom fall into universality classes that do not depend on the detailed structure of physics at arbitrarily high energies. If an interaction becomes stronger at large distances, we expect it to be relevant at low energies, whereas interactions that become weaker are irrelevant; anything else would be deemed unnatural. (The parallel between natural states being high entropy and natural theories arising from the renormalization group is an analogy rather than a rigorous equivalence, although there has been some tentative work towards establishing a more formal connection — in particular linking Boltzmann's H-theorem describing the evolution of entropy and the c-theorem describing renormalization-group flow^{2,3}. See for example ref. 4.)

If any system should be natural, it's the Universe. Nevertheless, according to the criteria just described, the Universe that we observe seems remarkably unnatural. The entropy of the Universe is not nearly as large as it could be, although it is at least increasing; for some reason, the early Universe was in a state of incredibly low entropy. And our fundamental theories of physics involve huge hierarchies between the energy scales characteristic of gravitation (the reduced Planck scale, $E_{\text{Pl}} = 1/\sqrt{8\pi G} \approx 10^{27}$ electron volts), particle physics (the Fermi scale of the weak interactions, $E_{\text{F}} \approx 10^{11}$ eV, and the scale of quantum chromodynamics, $E_{\text{QCD}} \approx 10^8$ eV, not to mention the ill-understood neutrino masses) and the recently

discovered vacuum energy ($E_{\text{vac}} \approx 10^{-3}$ eV). (Throughout this paper I will use units in which $\hbar = c = 1$, so that $1 \text{ eV} = 1.8 \times 10^{-33} \text{ g} = 5.1 \times 10^4 \text{ cm}^{-1} = 1.5 \times 10^{15} \text{ s}^{-1}$.) Table 1 includes some of these characteristic scales. Of course, it may simply be that the Universe is what it is, and these are brute facts that we have to live with. More optimistically, however, these apparently delicately tuned features of our Universe may be clues that can guide us towards a deeper understanding of the laws of nature.

Given this situation, physicists have been exploring dramatic extensions of our known theories, in an attempt to provide a larger context in which our apparently unnatural Universe is seen to make perfect sense. Interestingly, attempts to account for both the low entropy of the early Universe and the disparate energy scales of fundamental physics lead us to a similar idea: that the local Universe we observe is part of a much larger ensemble. This casts new light on the problems of naturalness, while raising vexing issues of its own; considerable advances in both theory and experiment will be necessary before we can decide whether we are learning the appropriate lessons from the clues provided by nature.

The state of our Universe

The state in which we find our observable Universe seems simple enough: on very large scales the distribution of matter is more or less homogeneous and isotropic, and distant galaxies are expanding away from each other in accordance with Hubble's law. But the early Universe was in fact in an extremely delicately specified state, analogous to a ball perched at the top of a hill. To account for this unusual initial condition, we need to ask what would really constitute a 'likely' state, and whether what we observe could arise from such a condition. If certain ideas about inflation and quantum gravity are correct, a universe like ours could come about as a quantum fluctuation from a pre-existing state that is simply empty space with a non-zero vacuum energy.

The Universe we see originated in a hot, dense state about 14 billion years ago. The deviations from perfect smoothness that we observe today have grown through gravitational instability from initially small perturbations of roughly equal amplitude on all length scales. Interestingly, the 'ordinary matter' made of particles described by the standard model of particle physics is only about 4% of the total energy of the Universe; another 23% is some 'dark matter' particle that has not yet been discovered. Even more interestingly, 73% of the Universe is 'dark energy', some form of smoothly distributed and nearly constant energy density^{5–7}. The most straightforward candidate for dark energy is vacuum energy, or the cosmological constant: an absolutely constant minimum energy of empty space itself, with a density $\rho_{\text{vac}} \approx E_{\text{vac}}^4 = (10^{-3} \text{ eV})^4$ (ref. 8).

If the Universe were in a likely configuration, we would expect it to be

¹Enrico Fermi Institute, Department of Physics and Kavli Institute for Cosmological Physics, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA.

in a high-entropy state: that is, in thermal equilibrium. Unfortunately, we do not have a rigorous definition of entropy for systems coupled to gravity, but we can make simple estimates. The entropy of matter and radiation (neglecting gravity) in our observable Universe is roughly the number of massless particles, $S_M(U) \approx 10^{88}$. This was the dominant contribution at early times when the matter distribution was essentially smooth; today, however, inhomogeneities have formed through gravitational collapse, creating large black holes at the centres of galaxies, and these black holes dominate the current entropy⁹. The Bekenstein–Hawking entropy of a black hole is proportional to its horizon area, $A: S_{BH} = A/4G \approx 10^{77} (M_{BH}/M_\odot)^2$, where $M_\odot$ is the mass of the Sun. As there are probably more than ten billion galaxies in the observable Universe with million-solar-mass black holes at their centres, the current entropy in black holes is at least $S_{BH}(U) \approx 10^{99}$. If we were to combine all of the matter in the observable Universe into one giant black hole, the entropy would be significantly larger, $S_{max} \approx 10^{120}$. So our Universe is currently in a rather low-entropy configuration, and it started out much lower. The fact that entropy tends to increase is of course just the second law of thermodynamics, and makes sense once we assume that the initial entropy was low. The puzzle is why the early Universe should differ so markedly from the late Universe, despite the intrinsic time-reversibility of the microscopic laws of physics; this is known as the ‘arrow of time’ problem.

Faced with the question of why our Universe started out with such a low entropy, most cosmologists would appeal to inflation^{10–13}. According to this idea, a very tiny patch of space dominated by temporary vacuum energy undergoes a period of rapidly accelerated expansion, smoothing out any inhomogeneities and ultimately reheating to the radiation-dominated early state of the conventional Big Bang model. For inflation to begin, the initial patch must itself be smooth¹⁴, but because it is so small, one imagines that it cannot be that hard to find an appropriate region somewhere in the chaotic conditions of the early Universe.

It is worth emphasizing that the only role of inflation is to explain the initial conditions of the observable Universe. And at this it does quite a good job: inflation predicts that the Universe should be spatially flat, and should have a scale-free spectrum of adiabatic density perturbations^{15–18}, both of which have been verified to respectable precision by observations of the cosmic microwave background⁷. But we are perfectly free to imagine that these features are simply part of the initial conditions—indeed, both spatial flatness and scale-free perturbations were investigated long before inflation. The only reason to invoke inflation is to provide a reason why such an initial condition would be natural.

However, as Penrose and others have argued, there is a skeleton in the inflationary closet, at least as far as entropy is concerned^{9,19,20}. The fact that the initial proto-inflationary patch must be smooth and dominated by dark energy implies that it must have a very low entropy itself; reasonable estimates for this entropy S_i range from about 1 to 10^{20} . Thus, among randomly chosen initial conditions, the likelihood of finding an appropriate proto-inflationary region is actually much less than simply finding the conditions of the conventional Big Bang model (or, for that matter, of our Universe ten minutes ago). It would seem that the conditions required to start inflation are less natural than those of the conventional Big Bang.

Table 1 | Orders of magnitude of the characteristic scales of our Universe, in units where $\hbar = c = 1$

Scale	Energy	Length
Planck (gravitation)	$E_{Pl} = (8\pi G)^{-1/2} \approx 10^{27}$ eV	$L_{Pl} \approx 10^{-32}$ cm
Fermi (weak interactions)	$E_F = (G_F)^{-1/2} \approx 10^{11}$ eV	$L_F \approx 10^{-16}$ cm
QCD (strong interactions)	$E_{QCD} = \Lambda_{QCD} \approx 10^8$ eV	$L_{QCD} \approx 10^{-13}$ cm
Vacuum (cosmological constant)	$E_{vac} = \rho_{vac}^{1/4} \approx 10^{-3}$ eV	$L_{vac} \approx 10^{-2}$ cm
Hubble (cosmology)	$E_H = H_0 \approx 10^{-33}$ eV	$L_H \approx 10^{28}$ cm

The reduced Planck energy is derived from Newton's constant G ; the Fermi scale of the weak interactions is derived from Fermi's constant G_F . The QCD scale Λ_{QCD} is the energy at which the strong coupling constant becomes large. The vacuum-energy scale arises from the energy density in the cosmological constant. The Hubble scale characteristic of cosmology is related to the total density ρ by the Friedmann equation, $H = \sqrt{(\rho)/E_{Pl}}$. Note the large dynamic range spanned by these parameters.

One possible escape from this conundrum has been suggested under the name of ‘spontaneous inflation’²¹, a particular implementation of the idea of ‘eternal inflation’^{22–25} (for related ideas see refs 26–28). In general relativity, high-entropy states correspond to empty space; given any configuration of matter, we can always increase the entropy by expanding the Universe and diluting the matter degrees of freedom²¹. But if empty space has a non-zero vacuum energy, it can be unstable to the creation of baby universes^{29–36}. In an empty universe with a positive cosmological constant (de Sitter space), quantum fluctuations will occasionally drive scalar fields to very large values of their potentials, setting up precisely the conditions necessary to begin inflation. The resulting bubble can branch off into a disconnected space-time, leading to an inflating region that would resemble our observable Universe.

The resulting ultra-large-scale structure of space-time is portrayed in Fig. 1. Time runs vertically in the empty background universe. Baby universes originate in quantum fluctuations, and break off from the background to form disconnected regions of space-time. In this way, each new universe naturally has very low entropy conditions at one boundary. The second law of thermodynamics is not violated, as the small region of the background from which the new universe arose has an even smaller entropy. The baby universes are portrayed as evolving sideways, because the locally defined arrow of time will not necessarily be smoothly connected to that of the background.

Needless to say, proposals of this type are extremely speculative, and may well be completely wrong. One mysterious aspect is the process of baby-universe creation, which certainly lies beyond the realm of established physics. In the context of string theory, there seem to be circumstances in which it can happen^{37–41} but also ones in which it cannot⁴²; whether it will occur in realistic situations is still unclear. Another issue is the nature of gravitational degrees of freedom in the context of the holographic principle, according to which degrees of freedom are distributed non-locally over volumes of space-time (see for example ref. 43). Given the relevance of inflation to observable phenomena, however, it is important to establish that inflation actually solves the problems it purports to address. In spontaneous inflation, there is a simple explanation for the low entropy of the initial state: it is not really ‘initial’, but rather arises by way of quantum fluctuations from a pre-existing de Sitter state with very large entropy and a very low entropy density. Whether this particular idea is on the right track or not, it is crucial to understand whether inflation plays a role in explaining how our observed configuration could be truly natural.

The laws of physics

The dynamical laws of nature at the microscopic level (including general relativity and the standard model of particle physics) are tightly constrained in the form that they may take, largely by symmetry principles such as gauge invariance and Lorentz invariance. The specific values of the numerical parameters of these theories are in principle arbitrary, although on naturalness grounds we would expect mass and energy scales to be roughly comparable to each other. As mentioned in the introduction, that is not what we observe: the Planck scale, Fermi scale, quantum chromodynamics (QCD) scale and vacuum-energy scale are separated by huge hierarchies. An intriguing possibility is that such hierarchies are not inevitable outcomes of deeper dynamics, but simply characterize the local conditions of our observable Universe, which is part of a much larger ensemble. In this case, what we think of as the ‘laws of physics’ are no more fundamental than the number of planets in our particular Solar System; they reflect environmental conditions rather than ineluctable truths.

We do (claim to) understand one of these hierarchies, that involving the QCD scale. For the strong interactions, the characteristic scale is governed by the logarithmic running of the QCD coupling constant, so a hierarchy is not that surprising. The difference between the Planck and Fermi scales ($E_{Pl}/E_F \approx 10^{16}$) is a celebrated puzzle in high-energy physics, known simply as ‘the hierarchy problem’; ideas such as supersymmetry may provide partial answers. The discrepancy between the Planck scale

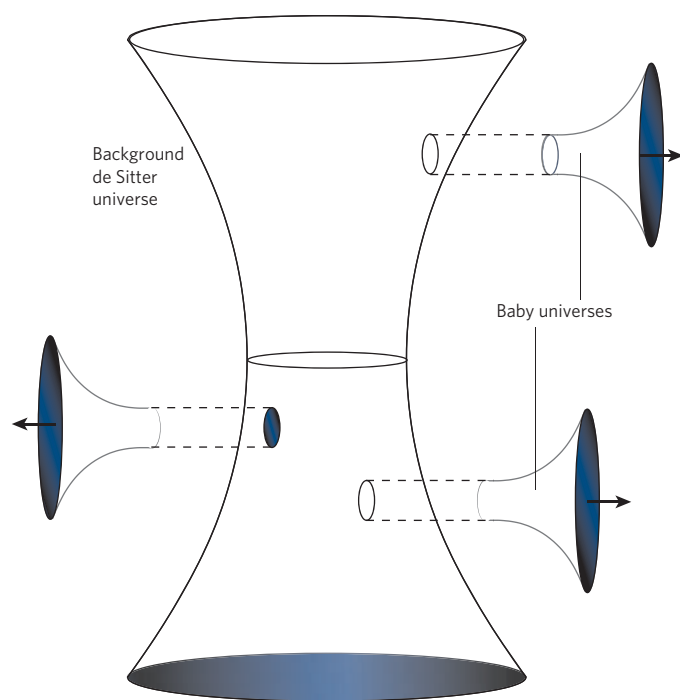


Figure 1 | A possible space-time diagram for the Universe on ultra-large scales. A natural state for a Universe with a positive vacuum energy is empty de Sitter space; this is the background universe in the figure, with time running vertically. In the presence of an appropriate scalar field, quantum fluctuations in such a background can lead to the nucleation of baby universes. Each baby universe is created in a proto-inflationary state, which then expands and reheats into a universe like that we observe. These universes are portrayed as evolving 'sideways', to emphasize that the local direction of time may not be related to that of the background space-time.

and the vacuum-energy scale ($E_{\text{pl}}/E_{\text{vac}} \approx 10^{30}$) is another puzzle, 'the cosmological constant problem'; most researchers agree that there are no compelling solutions currently on the market.

In contemplating the nature of these hierarchies, a complicating factor arises: we could not exist without them. If all of the energy scales of fundamental physics were roughly equal, it would be impossible to form structures in which quantum gravity did not play a significant role, and the large cosmological constant would make it difficult to imagine any complex structures at all (as the rapidly accelerating expansion of the Universe would work to tear things apart). One can imagine two very different attitudes in the face of this situation. (1) We just got lucky. The constants of nature just happen to take on values consistent with the existence of complex organisms. (2) Environmental selection. The parameters we observe are not truly fundamental, but merely reflect our local conditions; hospitable regions of the Universe will always show large hierarchies.

In the first case, there are two separate possibilities: either we are really lucky, in the sense that the observed hierarchies are truly unnatural and have no deeper explanation, or there exist unknown dynamical mechanisms that make these hierarchies perfectly natural. The latter possibility is obviously more attractive, although it is hard to tell whether such dynamical explanation will eventually be forthcoming. Environmental selection, sometimes discussed in terms of the 'anthropic principle', has received renewed attention since the discovery of the dark energy. The basic idea is undeniably true: if our observable Universe is only a tiny patch of a much larger 'multiverse' with a wide variety of local environments, there is a selection effect due to the fact that life can only arise in those regions that are hospitable to the existence of life. Of course, to give this tautology any explanatory relevance, it is necessary to imagine that such a multiverse exists.

Existence of the multiverse requires two conditions: the possibility of

multiple vacuum states with differing values of the constants of nature, and the realization of those states in distinct macroscopic regions. Recent ideas in string theory suggest that there may be a 'landscape' of metastable vacua arising from different ways of compacting extra dimensions in the presence of branes and gauge fields. Numbers such as 10^{500} vacua have been contemplated, promising more than enough diversity of possible local conditions^{44–49}. Meanwhile, the possibility of eternal inflation discussed in the previous section provides a mechanism for realizing such states: quantum fluctuations can lead to episodes of inflation that reheat into universe-sized domains in any of the permitted vacuum states⁵⁰. Thus, the ingredients of the multiverse picture seem to be in place, even if they remain speculative.

If all the multiverse does is allow for the existence of a region that resembles our own, it adds nothing to our understanding; it is equally sensible to say that our Universe is simply like that. Instead, the possible epistemological role of the multiverse is to explain why our observed parameters are natural. In principle, the multiverse picture allows us to predict the probability distribution for these parameters. In particular, the probability $P(X)$ that an observer measures their Universe to have feature X (such as "the ratio of the vacuum-energy scale to the Planck scale is of the order of 10^{-30} ") will be roughly of the form

$$P(X) = \frac{\sum_n \sigma_n(X) V_n \rho_n}{\sum_n V_n \rho_n} \quad (1)$$

In this expression, the index n labels all possible vacuum states; $\sigma_n(X)$ equals 1 if vacuum n has property X and 0 if it does not; V_n is the space-time volume in vacuum n ; and ρ_n is the space-time density of observers in vacuum n .

Obviously, there is considerable imprecision in the definition of equation (1); for example, we have been vague about what constitutes an 'observer', not to mention 'space-time volume' and 'density'. Interesting attempts have been made to formulate more precise versions of an equivalent expression^{51–56}. This expression suffices for our current purpose, however, which is to point out that actually calculating the probability is at best beyond our current abilities, and at worst completely hopeless. Just to mention the most obvious difficulty: in the context of eternal inflation, there is every reason to believe that the volumes V_n of some (if not all) vacua are infinite, and the expression is simply undefined.

The fact that equation (1) is undefined has not stopped people from trying to calculate it. The most famous example is Weinberg's prediction of the magnitude of the cosmological constant^{57–60}. This calculation imagines a flat prior distribution for the vacuum energy, keeping all other parameters fixed, and relies on the fact that the Universe recollapses in the presence of a large negative vacuum energy and expands too quickly for galaxies to form in the presence of a large positive vacuum energy. Under these assumptions, the predicted value of ρ_{vac} is not too different from what has been observed; moreover, the prediction was made before the measurement. Attempts have also been made to apply similar reasoning to models of particle physics^{61–65}.

Unfortunately, there is little reason to be satisfied with this calculation of the expected vacuum energy. In terms of equation (1), it is equivalent to imagining that the factor $\sigma_n(X)$ counting appropriate vacua is distributed uniformly in ρ_{vac} , the volume term V_n is simply a constant, and the density of observers ρ_n is proportional to the number of galaxies. The first of these is a (reasonable) guess, the second is likely to be fantastically wrong in the context of eternal inflation, and the last only makes sense if all of the other parameters are held fixed, which is not how we expect the multiverse to work. For example, allowing the amplitude of primordial density fluctuations to vary along with the vacuum energy can greatly change the result^{66–69}, not to mention variations of other parameters^{56,70}.

At present, then, there is no reliable environmental explanation for the observed value of the cosmological constant. To be fair, it does seem as though the allowed region in parameter space is very small, even if many parameters are allowed to vary, so it may be possible to make predictions about specific functions of the various parameters⁶⁸. Meanwhile, other attempts to use anthropic reasoning seem to lead to predictions that are in wild disagreement with observations⁷¹. But objections to the

credibility of the apparently successful predictions of the multiverse idea have equal force when applied to the apparently unsuccessful predictions; whether or not the idea is falsifiable, it would be an exaggeration to say that it has already been falsified.

More importantly, limitations in our current ability to calculate expectation values in the multiverse are not evidence that there is no truth to the idea itself. If we eventually decide that environmental selection plays no important role in explaining the observed parameters of nature, it will be because we come to believe that the parameters we measure locally are also characteristic of regions beyond our horizon, not because the very concept of the multiverse is aesthetically unacceptable or somehow a betrayal of the Enlightenment project of understanding nature through reason and evidence.

Future prospects

Naturalness is an ambiguous guide in the quest to understand our Universe better. The observation that a situation seems unnatural within a certain theoretical context does not carry anything like the force of an actual contradiction between theory and experiment. And despite our best efforts, naturalness is something that is hard to quantify objectively.

The search for naturalness plays an important role as a hint of physical processes that are not yet understood. In particle physics, attempts to find a natural solution to the hierarchy problem have driven investigations into supersymmetry and other models; in cosmology, attempts to explain the uniformity and flatness of our contemporary Universe helped to drive the development of inflationary cosmology. We can hope that attempts to understand the cosmological constant and the low entropy of the early Universe will lead to compelling new ideas about the fundamental architecture of nature.

The ideas discussed here involve the invocation of multiple inaccessible domains within an ultra-large-scale multiverse. For good reason, the reliance on the properties of unobservable regions and the difficulty in falsifying such ideas make scientists reluctant to grant them an explanatory role⁷². Of course, the idea that the properties of our observable domain can be uniquely extended beyond the cosmological horizon is an equally untestable assumption. The multiverse is not a theory; it is a consequence of certain theories (of quantum gravity and cosmology), and the hope is that these theories eventually prove to be testable in other ways. Every theory makes untestable predictions, but theories should be judged on the basis of the testable ones.

The ultimate goal is undoubtedly ambitious: to construct a theory that has definite consequences for the structure of the multiverse, such that this structure provides an explanation for how the observed features of our local domain can arise naturally, and that the same theory makes predictions that can be directly tested through laboratory experiments and astrophysical observations. To claim success in this programme, we will need to extend our theoretical understanding of cosmology and quantum gravity considerably, both to make testable predictions and to verify that some sort of multiverse picture really is a necessary consequence of these ideas. Only further investigation will allow us to tell whether such a programme represents laudable aspiration or misguided hubris.

1. 't Hooft, G. in *Recent Developments in Gauge Theories* vol. 59 (eds 't Hooft, G. et al.) NATO Advanced Study Institute, Series B: Physics (1980).
2. Zamolodchikov, A. B. 'Irreversibility' of the flux of the renormalization group in a 2-D field theory. *JETP Lett.* **43**, 730–732 (1986).
3. Cardy, J. L. Is there a c theorem in four-dimensions? *Phys. Lett. B* **215**, 749–752 (1988).
4. Gaiete, J. & O'Connor, D. Field theory entropy and the renormalization group. *Phys. Rev. D* **54**, 5163–5173 (1996).
5. Riess, A. G. et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116**, 1009–1038 (1998).
6. Perlmutter, S. et al. Measurements of Omega and Lambda from 42 high-redshift supernovae. *Astrophys. J.* **517**, 565–586 (1999).
7. Spergel, D. N. et al. First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: determination of cosmological parameters. *Astrophys. J.* **148** (suppl.), 175–194 (2003).
8. Carroll, S. M. The cosmological constant. *Living Rev. Rel.* **4**, 1 (2001).
9. Penrose, R. in *The Emperor's New Mind* Ch. 7 (Oxford Univ. Press, Oxford, 1989).
10. Guth, A. H. The inflationary universe: a possible solution to the horizon and flatness problems. *Phys. Rev. D* **23**, 347–356 (1981).
11. Linde, A. D. A new inflationary universe scenario: a possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. *Phys. Lett. B* **108**, 389–393 (1982).
12. Albrecht, A. & Steinhardt, P. J. Cosmology for grand unified theories with radiatively induced symmetry breaking. *Phys. Rev. Lett.* **48**, 1220–1223 (1982).
13. Linde, A. D. Chaotic inflation. *Phys. Lett. B* **129**, 177–181 (1983).
14. Vachaspati, T. & Trodden, M. Causality and cosmic inflation. *Phys. Rev. D* **61**, 023502 (2000).
15. Guth, A. H. & Pi, S. Y. Fluctuations in the new inflationary universe. *Phys. Rev. Lett.* **49**, 1110–1113 (1982).
16. Hawking, S. W. The development of irregularities in a single bubble inflationary universe. *Phys. Lett. B* **115**, 295–297 (1982).
17. Starobinsky, A. A. Dynamics of phase transition in the new inflationary universe scenario and generation of perturbations. *Phys. Lett. B* **117**, 175–178 (1982).
18. Bardeen, J. M., Steinhardt, P. J. & Turner, M. S. Spontaneous creation of almost scale-free density perturbations in an inflationary universe. *Phys. Rev. D* **28**, 679–693 (1983).
19. Hollands, S. & Wald, R. M. An alternative to inflation. *Gen. Rel. Grav.* **34**, 2043–2055 (2002).
20. Albrecht, A. & Sorbo, L. Can the Universe afford inflation? <arXiv:hep-th/0405270> (2004).
21. Carroll, S. M. & Chen, J. Spontaneous inflation and the origin of the arrow of time. <arXiv:hep-th/0410270> (2004).
22. Vilenkin, A. The birth of inflationary universes. *Phys. Rev. D* **27**, 2848–2855 (1983).
23. Linde, A. D. Eternal chaotic inflation. *Mod. Phys. Lett. A* **1**, 81–85 (1986).
24. Linde, A. D. Eternally existing selfreproducing chaotic inflationary universe. *Phys. Lett. B* **175**, 395–400 (1986).
25. Goncharov, A. S., Linde, A. D. & Mukhanov, V. F. The global structure of the inflationary universe. *Int. J. Mod. Phys. A* **2**, 561–591 (1987).
26. Garriga, J. & Vilenkin, A. Recycling universe. *Phys. Rev. D* **57**, 2230–2244 (1998).
27. Dutta, S. & Vachaspati, T. Islands in the Lambda-sea. *Phys. Rev. D* **71**, 083507 (2005).
28. Holman, R. & Mersini-Houghton, L. Why the universe started from a low entropy state. <arXiv:hep-th/0511102> (2005).
29. Farhi, E. & Guth, A. H. An obstacle to creating a universe in the laboratory. *Phys. Lett. B* **183**, 149–155 (1987).
30. Vilenkin, A. Quantum cosmology and the initial state of the Universe. *Phys. Rev. D* **37**, 888–897 (1988).
31. Farhi, E., Guth, A. H. & Guven, J. Is it possible to create a universe in the laboratory by quantum tunneling? *Nucl. Phys. B* **339**, 417–490 (1990).
32. Fischler, W., Morgan, D. & Polchinski, J. Quantum nucleation of false vacuum bubbles. *Phys. Rev. D* **41**, 2638–2641 (1990).
33. Fischler, W., Morgan, D. & Polchinski, J. Quantization of false vacuum bubbles: a Hamiltonian treatment of gravitational tunneling. *Phys. Rev. D* **42**, 4042–4055 (1990).
34. Linde, A. D. Hard art of the universe creation (stochastic approach to tunneling and baby universe formation). *Nucl. Phys. B* **372**, 421–442 (1992).
35. Aguirre, A. & Johnson, M. C. Dynamics and instability of false vacuum bubbles. *Phys. Rev. D* **72**, 103525 (2005).
36. Aguirre, A. & Johnson, M. C. Two tunnels to inflation. <arXiv:gr-qc/0512034> (2005).
37. Albergi, G. L., Lowe, D. A. & Trodden, M. Charged false vacuum bubbles and the AdS/CFT correspondence. *J. High Energy Phys.* **9907**, 020 (1999); published online 23 June 1999 (arXiv:hep-th/9906047).
38. Hellerman, S. On the landscape of superstring theory in $D > 10$. <arXiv:hep-th/0405041> (2005).
39. Adams, A., Liu, X., McGreevy, J., Saltman, A. & Silverstein, E. Things fall apart: topology change from winding tachyons. *J. High Energy Phys.* **0510**, 033 (2005); published online 5 August 2005 (arXiv:hep-th/0502021).
40. Dijkgraaf, R., Gopakumar, R., Ooguri, H. & Vafa, C. Baby universes in string theory. <arXiv:hep-th/0504221> (2005).
41. McGreevy, J. & Silverstein, E. The tachyon at the end of the universe. *J. High Energy Phys.* **0508**, 090 (2005); published online 3 August 2005 (arXiv:hep-th/0506130).
42. Freivogel, B. et al. Inflation in AdS/CFT. <arXiv:hep-th/0510046> (2005).
43. Dyson, L., Kleban, M. & Susskind, L. Disturbing implications of a cosmological constant. *J. High Energy Phys.* **0210**, 011 (2002); published online 14 November 2002 (arXiv:hep-th/0208013).
44. Bousso, R. & Polchinski, J. Quantization of four-form fluxes and dynamical neutralization of the cosmological constant. *J. High Energy Phys.* **0006**, 006 (2000); published online 26 June 2000 (arXiv:hep-th/0004134).
45. Feng, J. L., March-Russell, J., Sethi, S. & Wilczek, F. Saltatory relaxation of the cosmological constant. *Nucl. Phys. B* **602**, 307–328 (2001).
46. Giddings, S. B., Kachru, S. & Polchinski, J. Hierarchies from fluxes in string compactifications. *Phys. Rev. D* **66**, 106006 (2002).
47. Kachru, S., Kallosh, R., Linde, A. & Trivedi, S. P. De Sitter vacua in string theory. *Phys. Rev. D* **68**, 046005 (2003).
48. Douglas, M. R. The statistics of string/M theory vacua. *J. High Energy Phys.* **0305**, 046 (2003); published online 24 April 2003 (arXiv:hep-th/0303194).
49. Ashok, S. & Douglas, M. R. Counting flux vacua. *J. High Energy Phys.* **0401**, 060 (2004); published online 12 January 2004 (arXiv:hep-th/0307049).
50. Linde, A. Inflation, quantum cosmology and the anthropic principle. <arXiv:hep-th/0211048> (2002).
51. Vilenkin, A. Predictions from quantum cosmology. *Phys. Rev. Lett.* **74**, 846–849 (1995).
52. Tegmark, M. What does inflation really predict? *J. Cosmol. Astropart. Phys.* **0504**, 001 (2005); published online 12 October 2004 (arXiv:astro-ph/0410281).
53. Aguirre, A. On making predictions in a multiverse: conundrums, dangers, and coincidences. <arXiv:astro-ph/0506519> (2005).
54. Garriga, J., Schwartz-Perlov, D., Vilenkin, A. & Winitzki, S. Probabilities in the inflationary multiverse. <arXiv:hep-th/0509184> (2005).
55. Easther, R., Lim, E. A. & Martin, M. R. Counting pockets with world lines in eternal inflation. <arXiv:astro-ph/0511233> (2005).
56. Tegmark, M., Aguirre, A., Rees, M. & Wilczek, F. Dimensionless constants, cosmology and other dark matters. <arXiv:astro-ph/0511774> (2005).

57. Weinberg, S. Anthropic bound on the cosmological constant. *Phys. Rev. Lett.* **59**, 2607–2610 (1987).
58. Vilenkin, A. Quantum cosmology and the constants of nature. <arXiv:gr-qc/9512031> (1995).
59. Martel, H., Shapiro, P. R. & Weinberg, S. Likely values of the cosmological constant. *Astrophys. J.* **492**, 29–40 (1998).
60. Banks, T., Dine, M. & Motl, L. On anthropic solutions of the cosmological constant problem. *J. High Energy Phys.* **0101**, 031 (2001); published online 26 July 2000 (arXiv:hep-th/0007206).
61. Pogosian, L., Vilenkin, A. & Tegmark, M. Anthropic predictions for vacuum energy and neutrino masses. *J. Cosmol. Astropart. Phys.* **0407**, 005 (2004); published online 26 April 2004 (arXiv:astro-ph/0404497).
62. Arkani-Hamed, N. & Dimopoulos, S. Supersymmetric unification without low energy supersymmetry and signatures for fine-tuning at the LHC. *J. High Energy Phys.* **0506**, 073 (2005); published online 24 May 2004 (arXiv:hep-th/0405159).
63. Arkani-Hamed, N., Dimopoulos, S. & Kachru, S. Predictive landscapes and new physics at a TeV. <arXiv:hep-th/0501082> (2005).
64. Dine, M., O'Neil, D. & Sun, Z. Branches of the landscape. *J. High Energy Phys.* **0507**, 014 (2005); published online 20 January 2005 (arXiv:hep-th/0501214).
65. Fox, P. J. *et al.* Supersplit supersymmetry. <arXiv:hep-th/0503249> (2005).
66. Tegmark, M. & Rees, M. J. Why is the CMB fluctuation level 10^{-5} ? *Astrophys. J.* **499**, 526 (1998); published online 12 December 1997 (arXiv:astro-ph/9709058).
67. Graesser, M. L., Hsu, S. D. H., Jenkins, A. & Wise, M. B. Anthropic distribution for cosmological constant and primordial density perturbations. *Phys. Lett. B* **600**, 15–21 (2004).
68. Garriga, J. & Vilenkin, A. Anthropic prediction for Lambda and the Q catastrophe. <arXiv:hep-th/0508005> (2005).
69. Feldstein, B., Hall, L. J. & Watari, T. Density perturbations and the cosmological constant from inflationary landscapes. <arXiv:hep-th/0506235> (2005).
70. Aguirre, A. The cold big-bang cosmology as a counter-example to several anthropic arguments. *Phys. Rev. D* **64**, 083508 (2001); published online 12 July 2001 (arXiv:astro-ph/0106143).
71. Olum, K. D. Conflict between anthropic reasoning and observation. <arXiv:gr-qc/0303070> (2003).
72. Smolin, L. Scientific alternatives to the anthropic principle. <arXiv:hep-th/0407213> (2004).

Acknowledgements I thank A. Aguirre, J. Chen, B. Freivogel, J. Polchinski, E. Silverstein and A. Vilenkin for helpful conversations. This work was supported in part by the US Department of Energy, the National Science Foundation, and the David and Lucile Packard Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to the author (carroll@theory.uchicago.edu).

The large-scale structure of the Universe

Volker Springel¹, Carlos S. Frenk² & Simon D. M. White¹

Research over the past 25 years has led to the view that the rich tapestry of present-day cosmic structure arose during the first instants of creation, where weak ripples were imposed on the otherwise uniform and rapidly expanding primordial soup. Over 14 billion years of evolution, these ripples have been amplified to enormous proportions by gravitational forces, producing ever-growing concentrations of dark matter in which ordinary gases cool, condense and fragment to make galaxies. This process can be faithfully mimicked in large computer simulations, and tested by observations that probe the history of the Universe starting from just 400,000 years after the Big Bang.

The past two and a half decades have seen enormous advances in the study of cosmic structure, both in our knowledge of how it is manifest in the large-scale matter distribution, and in our understanding of its origin. A new generation of galaxy surveys — the 2-degree Field Galaxy Redshift Survey, or 2dFGRS¹, and the Sloan Digital Sky Survey, or SDSS² — have quantified the distribution of galaxies in the local Universe with a level of detail and on length scales that were unthinkable just a few years ago. Surveys of quasar absorption and of gravitational lensing have produced qualitatively new data on the distributions of diffuse intergalactic gas and of dark matter. At the same time, observations of the cosmic microwave background radiation, by showing us the Universe when it was only about 400,000 years old, have vindicated bold theoretical ideas put forward in the 1980s regarding the contents of the Universe and the mechanism that initially generated structure shortly after the Big Bang. The critical link between the early, near-uniform Universe and the rich structure seen at more recent times has been provided by direct numerical simulation. This has made use of the unremitting increase in the power of modern computers to create ever more realistic virtual universes: simulations of the growth of cosmic structure that show how astrophysical processes have produced galaxies and larger structures from the primordial soup. Together, these advances have led to the emergence of a ‘standard model of cosmology’ which, although seemingly implausible, has nevertheless been singularly successful.

Figure 1 strikingly illustrates how well this standard model can fit nearby structure. The observational wedge plots at the top and at the left show subregions of the SDSS and 2dFGRS, illustrating the large volume they cover in comparison to the ground-breaking Center for Astrophysics (CfA) galaxy redshift survey³ carried out during the 1980s (the central small wedge). These slices through the local three-dimensional galaxy distribution reveal a tremendous richness of structure. Galaxies, groups and clusters are linked together in a pattern of sheets and filaments that is commonly known as the ‘cosmic web’⁴. A handful of particularly prominent aggregations clearly stand out in these images, the largest containing of the order of 10,000 galaxies and extending for several hundred million light years. The corresponding wedge plots at the right and at the bottom show similarly constructed surveys of a virtual universe, the result of a simulation of the growth of structure and of the formation of galaxies in the current standard model of cosmology. The examples shown were chosen among a set of random ‘mock surveys’ to have large structures in similar positions to the real surveys. The similarity of structure between simulation and observation is striking,

and is supported by a quantitative comparison of clustering⁵. Here we review what we can learn from this excellent match.

The early 1980s produced two audacious ideas that transformed a speculative and notoriously uncertain subject into one of the most rapidly developing branches of physics. The first was the proposal that the ubiquitous dark matter that dominates large-scale gravitational forces consists of a new (and still unidentified) weakly interacting elementary particle. Because these particles are required to have small random velocities at early times, they were dubbed ‘cold dark matter’ or CDM. (Hot dark matter is also possible, for example a neutrino with a mass of a few tens of electron volts. Early cosmological simulations showed, however, that the galaxy distribution in a universe dominated by such particles would not resemble that observed⁶.) The second idea is ‘cosmic inflation’⁷, the proposal that the Universe grew exponentially for many doubling times perhaps $\sim 10^{-35}$ seconds after the Big Bang, driven by the vacuum energy density of an effective scalar field that rolls slowly from a false to the true vacuum. Quantum fluctuations in this ‘inflaton’ field are blown up to macroscopic scales and converted into genuine ripples in the cosmic energy density. These weak seed fluctuations grow under the influence of gravity and eventually produce galaxies and the cosmic web. Simple models of inflation predict the statistical properties of these primordial density fluctuations: their Fourier components should have random and independent phases and a near-scale-invariant power spectrum⁸. Inflation also predicts that the present Universe should have a flat geometry. With concrete proposals for the nature of the dark matter and for the initial fluctuation distribution, the growth of cosmic structure became, for the first time, a well-posed problem that could be tackled with the standard tools of physics.

The backbone of the cosmic web is the clumpy yet filamentary distribution of dark matter. The presence of dark matter was first inferred from the dynamics of galaxy clusters by Zwicky⁹. But it took over half a century for dark matter to become an integral part of our view of galaxies and of the Universe as a whole, and for its average density to be estimated reliably. Today, the evidence for the pervasive presence of dark matter is overwhelming and includes galactic rotation curves, the structure of galaxy groups and clusters, large-scale cosmic flows and, perhaps most directly, gravitational lensing, a phenomenon first proposed as an astronomical tool by Zwicky himself¹⁰. The distorted images of background galaxies as their light travels near mass concentrations reveal the presence of dark matter in the outer haloes of galaxies^{11,12}, in galaxy clusters¹³ and in the general mass field¹⁴.

¹Max-Planck-Institute for Astrophysics, Karl-Schwarzschild-Strasse 1, 85740 Garching, Germany. ²Institute for Computational Cosmology, Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK.

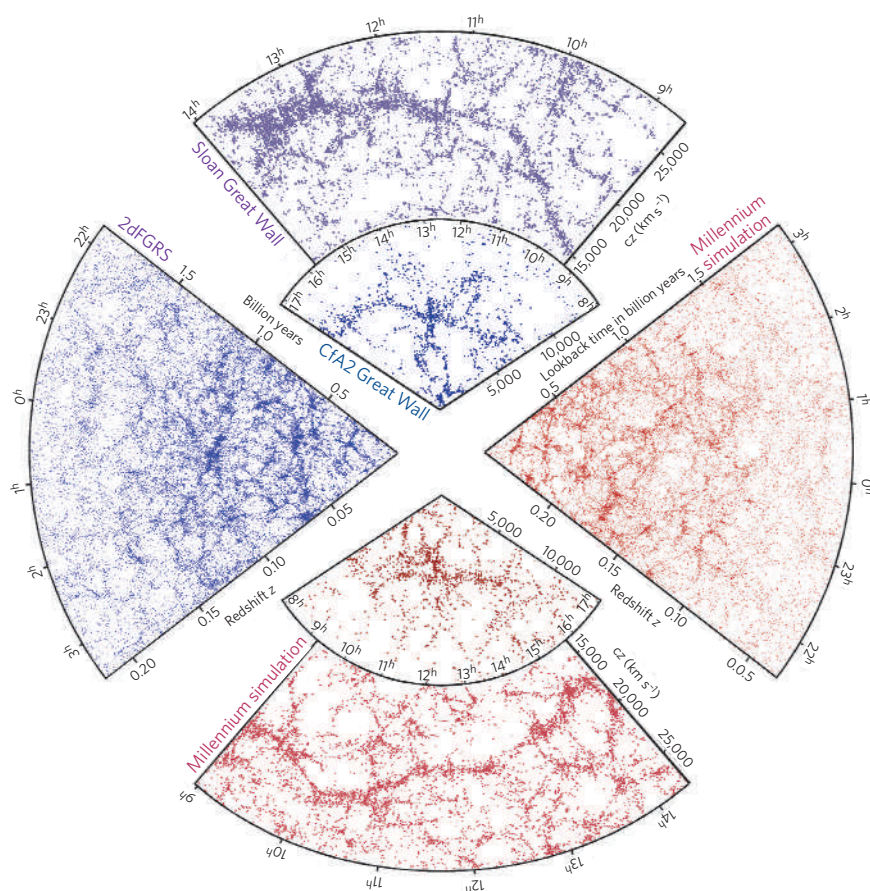


Figure 1 | The galaxy distribution obtained from spectroscopic redshift surveys and from mock catalogues constructed from cosmological simulations. The small slice at the top shows the CfA2 'Great Wall'²³, with the Coma cluster at the centre. Drawn to the same scale is a small section of the SDSS, in which an even larger 'Sloan Great Wall' has been identified¹⁰⁰. This is one of the largest observed structures in the Universe, containing over 10,000 galaxies and stretching over more than 1.37 billion light years. The cone on the left shows one-half of the 2dFGRS, which determined distances to more than 220,000 galaxies in the southern sky out to a depth of 2 billion light years. The SDSS has a similar depth but a larger solid angle and currently includes over 650,000 observed redshifts in the northern sky. At the bottom and on the right, mock galaxy surveys constructed using semi-analytic techniques to simulate the formation and evolution of galaxies within the evolving dark matter distribution of the 'Millennium' simulation⁵ are shown, selected with matching survey geometries and magnitude limits.

When expressed in units of the critical density required for a flat cosmic geometry, the mean density of dark matter is usually denoted by Ω_{dm} . Although a variety of dynamical tests have been used to constrain Ω_{dm} , in general such tests give ambiguous results because velocities are induced by the unseen dark matter and the relation of its distribution to that of the visible tracers of structure is uncertain. The notion of a substantial bias in the galaxy distribution relative to that of dark matter was introduced in the 1980s to account for the fact that different samples of galaxies or clusters are not directly tracing the underlying matter distribution^{15–17}. Defined simply as the ratio of the clustering strengths, the 'bias function' was also invoked to reconcile low dynamical estimates for the mass-to-light ratio of clusters with the high global value required in the theoretically preferred flat, $\Omega_{\text{dm}} = 1$ universe. But because massive clusters must contain approximately the universal mix of dark matter and baryons (ordinary matter), this uncertainty is neatly bypassed by comparing the measured baryon fraction in clusters with the universal fraction under the assumption that the mean baryon density, Ω_{b} , is the value inferred from Big Bang nucleosynthesis¹⁸. Applied to the Coma cluster, this simple argument gave $\Omega_{\text{dm}} \leq 0.3$ where the inequality arises because some or all of the dark matter could be baryonic¹⁸. This was the first determination of $\Omega_{\text{dm}} < 1$ that could not be explained away by invoking bias. Subsequent measurements have confirmed the result¹⁹ which also agrees with recent independent estimates based, for example, on the relatively slow evolution of the abundance of galaxy clusters^{20,21} or on the detailed structure of fluctuations in the microwave background radiation²².

The mean baryon density implied by matching Big Bang nucleosynthesis to the observed abundances of the light elements is only $\Omega_{\text{b}} h^2 \approx 0.02$, where h denotes the Hubble constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Dynamical estimates, although subject to bias uncertainties, have long suggested that $\Omega_{\text{m}} = \Omega_{\text{dm}} + \Omega_{\text{b}} \approx 0.3$, implying that the dark matter cannot be baryonic. Plausibly it is made up of the hypothetical elementary particles postulated in the 1980s, for example axions or the lowest mass supersymmetric partner of the known particles. Such

low estimates of the mean matter density Ω_{m} are incompatible with the flat geometry predicted by inflation unless the Universe contains an additional unclustered and dominant contribution to its energy density, for example a cosmological constant Λ such that $\Omega_{\text{m}} + \Omega_{\Lambda} \approx 1$. Two large-scale structure surveys carried out in the late 1980s, the APM (automated photographic measuring) photographic survey²³ and the QDOT redshift survey of infrared galaxies²⁴, showed that the power spectrum of the galaxy distribution, if it traces that of the mass on large scales, can be fitted by a simple CDM model only if the matter density is low, $\Omega_{\text{m}} \approx 0.3$. This independent confirmation of the dynamical arguments led many to adopt the now standard model of cosmology, Λ CDM.

It was therefore with a mixture of amazement and *déjà vu* that cosmologists greeted the discovery in 1998 of an accelerated cosmic expansion^{25,26}. Two independent teams used distant type Ia supernovae to perform a classical observational test. These 'standard candles' can be observed out to redshifts beyond 1. Those at $z \geq 0.5$ are fainter than expected, apparently indicating that the cosmic expansion is currently speeding up. Within the standard Friedmann cosmology, there is only one agent that can produce an accelerating expansion: the cosmological constant first introduced by Einstein, or its possibly time- or space-dependent generalization, 'dark energy'. The supernova evidence is consistent with $\Omega_{\Lambda} \approx 0.7$, just the value required for the flat universe predicted by inflation.

The other key prediction of inflation, a density fluctuation field consistent with amplified quantum noise, received empirical support from the discovery by the COsmic Background Explorer (COBE) satellite in 1992 of small fluctuations in the temperature of the cosmic microwave background (CMB) radiation²⁷. These reflect primordial density fluctuations, modified by damping processes in the early Universe which depend on the matter and radiation content of the Universe. More recent measurements of the CMB^{28–32} culminating with those by the WMAP (Wilkinson Microwave Anisotropy Probe) satellite²² have provided a striking confirmation of the inflationary CDM model: the measured temperature fluctuation spectrum is nearly scale-invariant on large

scales and has a series of ‘acoustic’ peaks that reflect the coherent oscillations experienced by the photon–baryon fluid before the moment when the primordial plasma recombined and the radiation escaped. The fluctuation spectrum depends on the parameters that define the geometry and content of the Universe and the initial fluctuation distribution, so their values are constrained by the data. In practice, there are degeneracies among the parameters, and the strongest constraints come from combining the CMB data with other large-scale structure datasets. Present estimates^{22,33–36} give a flat universe with $\Omega_{\text{dm}} = 0.20 \pm 0.020$, $\Omega_{\text{b}} = 0.042 \pm 0.002$, $\Omega_{\Lambda} = 0.76 \pm 0.020$, $h = 0.74 \pm 0.02$. The consistency of these values with other independent determinations and the close agreement of the CMB data with theoretical predictions formulated over 20 years earlier³⁷ belong amongst the most remarkable successes of modern cosmology.

The growth of large-scale structure

The microwave background radiation provides a clear picture of the young Universe, where weak ripples on an otherwise uniform sea display a pattern that convincingly supports our standard model for the cosmic mass/energy budget and for the process that initially imprinted cosmic structure. At that time there were no planets, no stars, no galaxies, none of the striking large-scale structures seen in Fig. 1. The richness of the observed astronomical world grew later in a complex and highly nonlinear process driven primarily by gravity. This evolution can be followed in detail only by direct numerical simulation. Early simulations were able to reproduce qualitatively the structure observed both in large galaxy surveys and in the intergalactic medium^{16,38}. They motivated the widespread adoption of the CDM model well before it gained support from microwave background observations. Many physical processes affect galaxy formation, however, and many aspects must be treated schematically within even the largest simulations. The resulting uncertainties are best estimated by exploring a wide range of plausible descriptions and checking results against observations of many different types. The main contribution of early CDM galaxy formation modelling was perhaps the dethroning of the ‘island universe’ or ‘monolithic collapse’ paradigm and the realization that galaxy formation is a process extending from early times to the present day, rather than an event that occurred in the distant past³⁹.

In a Λ CDM universe, quasi-equilibrium dark matter clumps or ‘haloes’ grow by the collapse and hierarchical aggregation of ever more massive systems, a process described surprisingly well by the phenomenological model of Press and Schechter and its extensions^{40,41}. Galaxies form at the centres of these dark haloes by the cooling and condensation of gas which fragments into stars once it becomes sufficiently dense⁴². Groups and clusters of galaxies form as haloes aggregate into larger systems. They are arranged in the ‘cosmic web’, the larger-scale pattern of filaments and sheets which is a nonlinear gravitational ‘sharpening’ of

the pattern already present in the gaussian random field of initial fluctuations⁴. The first observable objects were probably massive stars collapsing in small haloes and switching on at redshifts of 50 and higher⁴³. By a redshift of 15 these may have been sufficiently numerous for their radiation to re-ionize all the gas in the Universe⁴⁴. So far they have not been observed directly, but it is one of the main goals of the next generation of low-frequency radio telescopes to observe their effects directly in the strongly redshifted 21-cm transition of neutral hydrogen.

Detailed simulations from Λ CDM initial conditions have been used to study the formation of the first luminous objects and the re-ionization of the Universe, but these still await testing against observation^{44,45}. In contrast, predictions for the structure, the ionization state and the heavy element content of intergalactic gas at redshifts below 6 can be checked in detail against absorption features observed in the spectra of distant quasars which provide, in effect, a one-dimensional topographic image of the intervening large-scale structure.

As an example, Fig. 2 shows a typical high-resolution spectrum of a distant quasar at redshift $z = 3.26$. At shorter wavelengths than the Lyman α emission line of the quasar, there is a ‘forest’ of absorption lines of differing strength. The modern interpretation is that these features arise from Lyman α absorption by the smoothly varying distribution of foreground intergalactic hydrogen, in effect from the filaments, sheets and haloes of cosmic structure. It was a conceptual breakthrough, and an important success for the CDM paradigm, when hydrodynamical simulations showed that this interpretation could explain in detail the observed statistics of the absorption lines^{38,46}. Considerable recent advances both in the quality and in the quantity of data available have made it possible to measure a variety of statistics for the Lyman α forest as a function of redshift to high precision^{47–49}. Comparing with appropriately designed numerical simulations has provided strong confirmation of the underlying paradigm at a level that is remarkable, given the evidence that intergalactic gas is contaminated with galaxy ejecta in a way that the simulations do not yet adequately reproduce^{36,50–52}. This approach has also helped to strengthen constraints on the paradigm’s parameters, in particular on the spectrum of fluctuations produced by inflation and on the masses of neutrinos.

At lower redshift direct and quantitative measures of large-scale structure can be obtained from the weak, coherent distortions of the images of faint galaxies induced by gravitational lensing as their light travels through the intervening cosmic web⁵³. The distortions depend only on the gravitational field in intergalactic space and so lensing data test predictions for the mass distribution in a way that is almost independent of the complex astrophysics that determines the observable properties of galaxies. The lensing effect is very weak, but can be measured statistically to high precision with large enough galaxy samples.

As an example, Fig. 3 shows a measure of the mean square coherent distortion of distant galaxy images within randomly placed circles on the

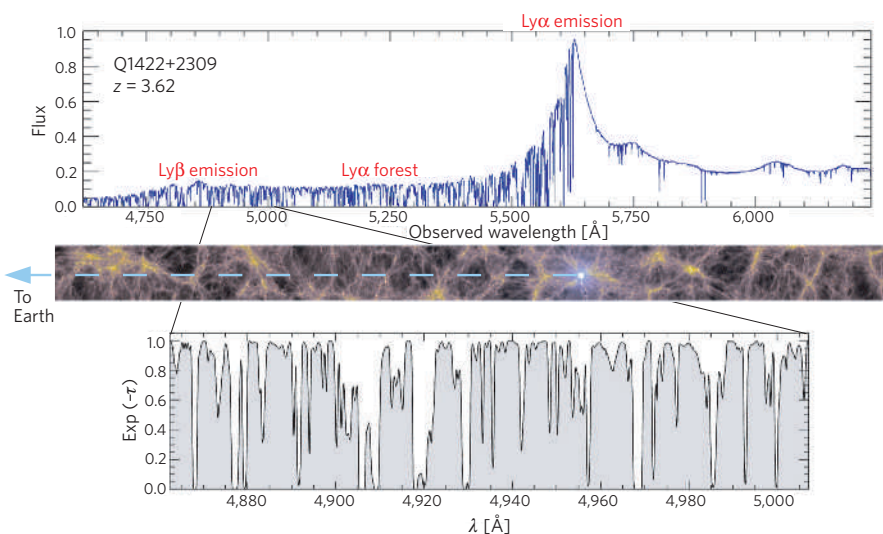


Figure 2 The Lyman α forest as a probe of large-scale structure. The panel on the top shows a typical high-resolution spectrum of a quasar at redshift $z = 3.62$. Shortward of the redshifted Lyman α emission line at $1216(1+z)$ Å, the spectrum shows a ‘forest’ of absorption lines of different strength produced by intervening neutral hydrogen gas along the line-of-sight from the quasar to the Earth. Hydrodynamical simulations reproduce the observed absorption spectra with remarkable fidelity, as illustrated by the simulated spectrum in the bottom panel, corresponding to intervening large-scale structure at $z \approx 3$. The sketch in the middle panel shows an example of the gas distribution in a simulated Λ CDM model.

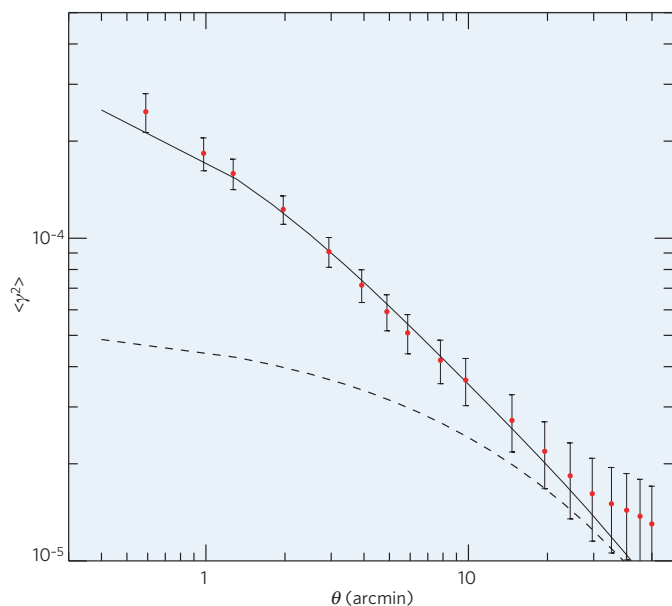


Figure 3 | Variance of the weak lensing shear as a function of top-hat smoothing scale. The data points show recent measurements from the VIRMOS survey⁵⁴. The solid line gives the predicted signal for the nonlinear mass distribution in the standard Λ CDM model (normalized so that the linear mass overdensity in spheres of radius $8 h^{-1}$ Mpc is $\sigma_8 = 0.84$), and the dashed line shows a linear extrapolation based on the structure present at early times. Because the weak lensing shear depends sensitively on the nonlinear clustering of the total mass distribution, it provides a particularly powerful probe of cosmology. Figure courtesy of Ludo van Waerbeke.

sky as a function of the radius of those circles⁵⁴. Clearly, the distortion is detected with very high significance. The two curves show the predicted signal in the standard Λ CDM model based on (i) detailed simulations of the growth of structure in the dark matter distribution, and (ii) a simple linear extrapolation from the structure present at early times. Nonlinear effects are strong because the distortions are dominated by the gravity of individual dark matter haloes. Meaningful comparison between theory and observation thus requires high-precision large-scale structure simulations, and generating these constitutes a great numerical challenge. Similar lensing measurements, but now within circles centred on observed galaxies (rather than random points), can be used to determine the average total mass surrounding galaxies as a function of radius, redshift and galaxy properties⁵⁵. This wealth of information can only be interpreted by simulations that follow both the dark matter distribution and the formation and evolution of the galaxy population.

The Lyman α forest and gravitational lensing thus provide windows onto the large-scale structure of the Universe that complement those obtained from galaxy surveys by extending the accessible redshift range and, more importantly, by measuring the structure in the diffuse gas and in the total mass distribution rather than in the distribution of galaxies. In principle, these measures should have different (and perhaps weaker) sensitivity to the many uncertain aspects of how galaxies form. Remarkably, all three measures are consistent both with each other and with the standard model at the level that quantitative comparison is currently possible^{36,54,56}.

Galaxy surveys such as those illustrated in Fig. 1 contain an enormous amount of information about large-scale structure. The strength of clustering is known to depend not only on galaxy luminosity, colour, morphology, gas content, star-formation activity, type and strength of nuclear activity and halo mass, but also on the spatial scale considered and on redshift. Such dependences reflect relations between the formation histories of galaxies and their larger-scale environment. Some (for example, the dependence on halo or galaxy mass) are best thought of as deriving from the statistics of the initial conditions. Others (for example the dependence on nuclear or star-formation activity) seem more naturally associated with late-time environmental influences. Early studies

attempted to describe the relation between the galaxy and mass distributions by a bias function. Recent data suggest that this concept is of limited value, except, perhaps, on the largest scales; bias estimates depend not only on scale, redshift and galaxy properties, but also on the particular measure of clustering studied. Understanding the link between the mass and galaxy distributions requires realistic simulations of the galaxy formation process throughout large and representative regions of the Universe. Given the complexity of galaxy formation, such simulations must be tuned 'by hand' to match as many of the observed properties of galaxies as possible. Only if clustering turns out to be insensitive to such tuning can we consider the portrayal of large-scale structure to be robust and realistic.

In Fig. 4, we show the time evolution of the mass and galaxy distributions in a small subregion of the largest simulation of this type yet⁵. The emergence of the cosmic web can be followed in stunning detail, producing a tight network of filaments and walls surrounding a foam of voids. This characteristic morphology was seen in the first generation of cold dark matter simulations carried out over 20 years ago¹⁶, but the match was not perfect; the recipe adopted to relate the galaxy and mass distributions was too crude to reproduce in detail the clustering of galaxies. It has taken models like those of Fig. 4 to explain why the observed galaxy autocorrelation function is close to a power law whereas the simulated dark matter autocorrelation function shows significant features^{5,57}.

Simulated autocorrelation functions for dark matter and for galaxies are shown in Fig. 5 for the same times imaged in Fig. 4. The shape difference between the two is very evident, and it is remarkable that at $z = 0$ the power-law behaviour of the galaxy correlations extends all the way down to 10 kpc, the observed size of galaxies. Similar behaviour has recently been found for luminous red galaxies in the Sloan Digital Sky Survey⁵⁸. The galaxy distribution in this simulation also reproduces the observed dependence of present-day clustering on luminosity and colour⁵ as well as the observed galaxy luminosity functions, the observationally inferred formation histories of elliptical galaxies, and the bimodal colour-magnitude distribution observed for galaxies^{59,60}.

A striking feature of Fig. 4 is the fact that while the growth of large-scale structure is very clear in the mass distribution, the galaxy distributions appear strongly clustered at all times. This difference shows up dramatically in the autocorrelation functions plotted in Fig. 5 and has been a prediction of CDM theories since the first simulations including crude bias recipes¹⁶. A decade later when direct measurements of galaxy clustering at redshifts as high as $z \approx 3$ –4 found "surprisingly" large amplitudes, comparable to those found in the present-day Universe^{61,62}, the results turned out to be in good agreement with estimates based on more detailed modelling of galaxy formation in a CDM universe^{63,64}. In effect, the galaxies already outline the pattern of the cosmic web at early times, and this pattern changes relatively little with the growth of structure in the underlying dark matter distribution.

Could the standard model be wrong?

Given the broad success of the Λ CDM model, is it conceivable that it might be wrong in a significant way requiring a fundamental revision? The concordance of experimental results relying on a variety of physical effects and observed over a wide range of cosmic epochs suggests that this is unlikely. Nevertheless, it is clear that some of the most fundamental questions of cosmology (what is the dark matter? the dark energy?) remain unanswered. In addition, some of the key observational underpinnings of the model still carry worrying uncertainties. Can we use our ever-improving measurements of large-scale structure to carry out critical tests?

Perhaps the deepest reason to be suspicious of the paradigm is the apparent presence of a dark energy field that contributes $\sim 70\%$ of the Universe's content and has, for the past 5 billion years or so, driven an accelerated cosmic expansion. Dark energy is problematic from a field theoretical point of view⁶⁵. The simplest scenario would ascribe a vacuum energy to quantum loop corrections at the Planck scale, hc^5/G , which is of the order of 10^{19} GeV, where gravity should unify with the other fundamental forces. This is more than 120 orders of magnitude

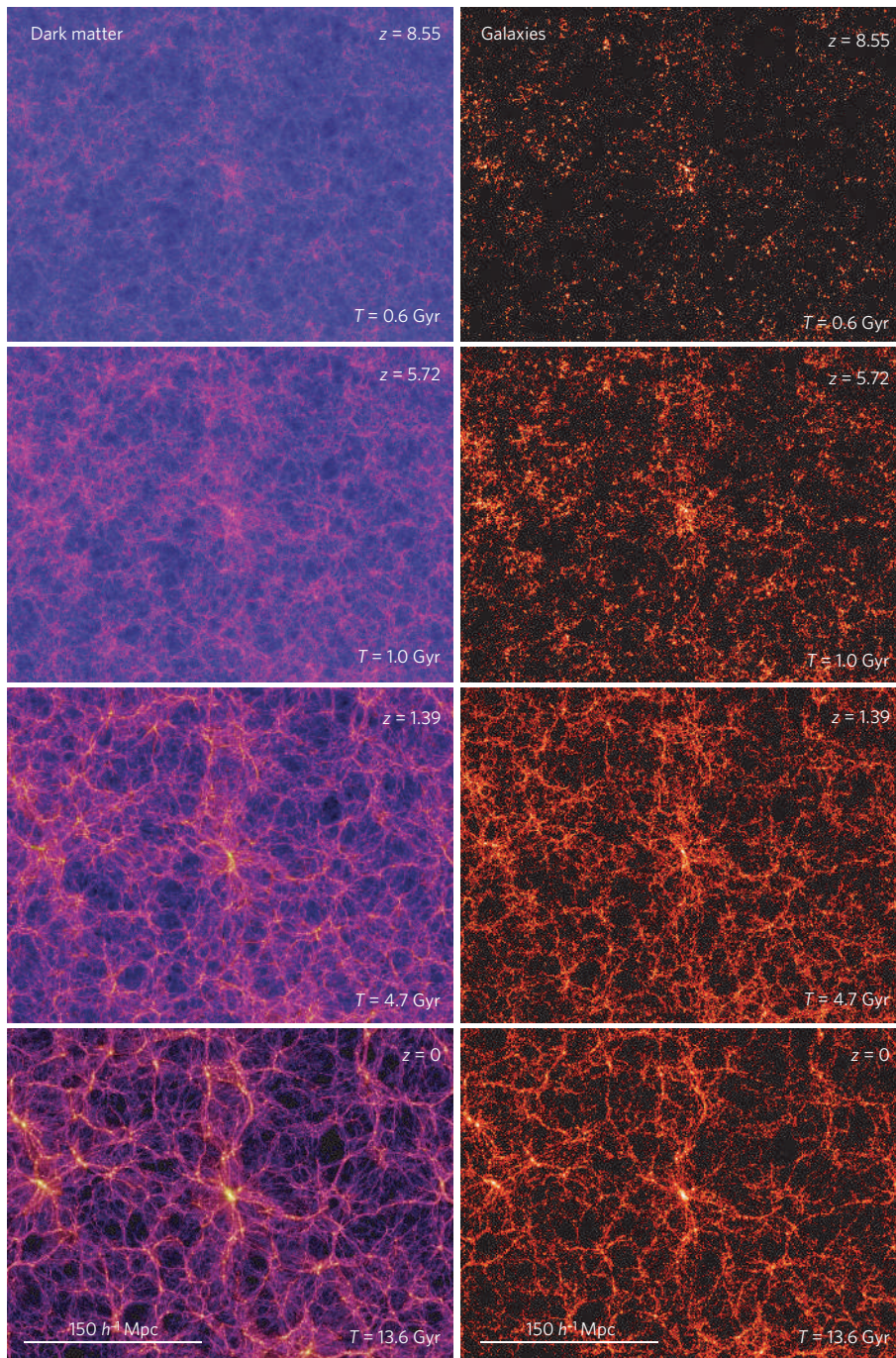


Figure 4 | Time evolution of the cosmic large-scale structure in dark matter and galaxies, obtained from cosmological simulations of the Λ CDM model. The panels on the left show the projected dark matter distribution in slices of thickness $15 h^{-1}$ Mpc, extracted at redshifts $z = 8.55$, $z = 5.72$, $z = 1.39$ and $z = 0$ from the Millennium N -body simulation of structure formation⁵. These epochs correspond to times of 600 million, 1 billion, 4.7 billion and 13.6 billion years after the Big Bang, respectively. The colour hue from blue to red encodes the local velocity dispersion in the dark matter, and the brightness of each pixel is a logarithmic measure of the projected density. The panels on the right show the predicted distribution of galaxies in the same region at the corresponding times obtained by applying semi-analytic techniques to simulate galaxy formation in the Millennium simulation⁵. Each galaxy is weighted by its stellar mass, and the colour scale of the images is proportional to the logarithm of the projected total stellar mass. The dark matter evolves from a smooth, nearly uniform distribution into a highly clustered state, quite unlike the galaxies, which are strongly clustered from the start.

larger than the value required by cosmology. Postulating instead a connection to the energy scale of quantum chromodynamics would still leave a discrepancy of some 40 orders of magnitude. A cosmological dark energy field that is so unnaturally small compared with these particle physics scales is a profound mystery.

The evidence for an accelerating universe provided by type Ia supernovae relies on a purely phenomenological calibration of the relation between the peak luminosity and the shape of the light curve. It is this that lets these supernovae be used as an accurate standard candle. Yet this relation is not at all understood theoretically. Modern simulations of thermonuclear explosions of white dwarfs suggest that the peak luminosity should depend on the metallicity of the progenitor star^{66,67}. This could, in principle, introduce redshift-dependent systematic effects, which are not well constrained at present. Perhaps of equal concern is the observation that the decline rate of type Ia supernovae correlates with host galaxy type^{68,69}, in the sense that the more luminous supernovae (which decline more slowly) are preferentially found in spiral galaxies.

Interestingly, it has also been pointed out that without the evidence for accelerated expansion from type Ia supernovae, a critical density Einstein–de Sitter universe can give a good account of observations of large-scale structure provided the assumption of a single power law for the initial inflationary fluctuation spectrum is dropped, a small amount of hot dark matter is added, and the Hubble parameter is dropped to the perhaps implausibly low value $h \approx 0.45$ (ref. 70).

The CMB temperature measurements provide particularly compelling support for the paradigm. The WMAP temperature maps do, however, show puzzling anomalies that are not expected from gaussian fluctuations^{71–73}, as well as large-scale asymmetries that are equally unexpected in an isotropic and homogeneous space^{74,75}. Although these signals could perhaps originate from foregrounds or residual systematics, it is curious that the anomalies seem well matched by anisotropic Bianchi cosmological models, although the models examined so far require unacceptable cosmological parameter values⁷⁶. Further data releases from WMAP and future CMB missions such as PLANCK will shed light on these

peculiarities of the current datasets. Perhaps the anomalous effects will go away; or they could be the first signs that the standard model needs substantial revision.

The unknown nature of the dark matter is another source of concern. Is the dark matter really 'cold' and non-interacting, and is it really dark? Does it exist at all? Until the posited elementary particles are discovered, we will not have definitive answers to these questions. Already there are hints of more complicated possibilities. It has been suggested, for instance, that the γ -ray excess flux recently detected in the direction of the Galactic Centre⁷⁷ might be due to self-annihilating dark matter particles⁷⁸, an idea that is, in principle, plausible for a range of dark matter candidates in supersymmetric field theories. Alternative theories of gravity, most notably modified newtonian dynamics (MOND)⁷⁹ have been proposed to do away with the need for dark matter altogether. Although MOND can explain the rotation curves of galaxies, on other scales the theory does not seem to fare so well. For example, although it can account for the total mass in galaxy clusters, MOND requires the presence of large amounts of unseen material within the central few kiloparsecs of the cluster cores⁸⁰. It has yet to be demonstrated convincingly that MOND can reproduce observed large-scale structure starting from the initial conditions imaged in the CMB and so pass the test illustrated in Fig. 1.

At present the strongest challenge to Λ CDM arises not from large-scale structure, but from the small-scale structure within individual galaxies. It is a real possibility that the model could be falsified by measurements of the distribution and kinematics of matter within galaxies, and some astronomers argue that this has, in fact, already happened. The internal structure of dark matter haloes predicted by the Λ CDM model can be calculated quite precisely from high-resolution simulations. These predict the survival of a large number of self-bound substructures which orbit within haloes^{81,82}, as well as a universal halo density profile which is cusped in the middle, corresponding to a steeply rising rotation curve⁸³. Unfortunately, the effects of galaxy formation within a dark matter halo are difficult to calculate, accounting, in part, for the lively debate that continues to rage over whether the measured rotation curves of dwarf and low surface brightness galaxies are in conflict with the theory^{84,85}. The second contentious issue on galaxy scales, the small number of observed satellites, may have been resolved by identifying astrophysical processes that could have rendered most of the surviving

subhaloes invisible^{86,87}. Gravitational lensing measurements may offer a test of this explanation⁸⁸. Lensing also allows independent determinations of halo density profiles, a method that has in fact led to new challenges for Λ CDM. Recent results on cluster scales favour steeper inner mass profiles than expected, but the significance of this discrepancy is unclear because of uncertainties originating in halo triaxiality and projection effects⁸⁹.

Future tests of large-scale structure and cosmology

Very few of the important questions in cosmology and large-scale structure can be regarded as closed. The recent history of the subject provides a vivid reminder of how new theoretical insights and/or new observational datasets can quickly overturn conventional wisdom in rapidly advancing fields of science. At the present time, the two outstanding questions are the identity of the dark matter and the nature of the dark energy.

There is every reason to be optimistic about the prospects of detecting cold dark matter particles from the halo of our Galaxy, either directly in laboratory searches or indirectly through particle annihilation radiation. Additionally, if cold dark matter is indeed a supersymmetric particle, evidence for its existence may be forthcoming from experiments at CERN's large-hadron collider⁹⁰.

Unravelling the nature of the dark energy is a much more daunting task. A strategy that has gained momentum in recent years is to set tighter empirical constraints on the amount of dark energy and on its possible time evolution. Large projects such as the Joint Dark Energy Mission, currently at an early design phase by NASA, are being planned to measure the equation of state parameter, $w = P/(\rho c^2)$, of the dark energy, where P is the 'dark pressure' of the vacuum, and its time evolution, $w' = dw/dz$. The hope is that such empirical constraints will clarify the nature of the dark energy and perhaps point to a field-theoretical explanation. The range of possibilities is large. We might find that the dark energy interacts with the dark matter, or that the dark energy is not a field at all but rather a manifestation of some nonlinear effect within general relativity or one of its extensions.

Progress towards constraining dark energy is likely to come both from refinements of classical cosmological probes and from entirely new ways to study large-scale structure. Examples in the first category include measuring the abundance of galaxy clusters as a function of cosmic time. This probes the growth of the mass fluctuation spectrum and the variation of the cosmological volume element⁹¹. Extending such measurements to redshifts $z \geq 1$ may set useful constraints on the dark energy equation of state, provided systematic effects can be kept under control. Also promising are observations of high-redshift type Ia supernovae for much larger samples than have been accumulated so far. Again, it will be crucial to control systematic effects. The PLANCK satellite mission and subsequent polarization-optimized experiments will make definitive measurements of the CMB and perhaps unlock some of its last secrets.

Examples of new tests of the large-scale structure include weak lensing tomography and the study of baryon oscillations in the matter distribution at late times. The physical mechanism that generated acoustic peaks in the CMB temperature power spectrum also imprinted an oscillatory feature in the linear power spectrum of the dark matter⁹². The Virgo consortium's Millennium simulation, illustrated in Fig. 1 and Fig. 4, demonstrated that the oscillations survive the destructive influence of nonlinear gravitational evolution even to the present day, albeit in distorted form⁵. Most importantly, this simulation also demonstrated that these 'baryon wiggles' should be visible in suitably selected galaxy samples. Early indications suggest that the baryon oscillations in the galaxy distribution have, in fact, been detected in the 2dFGRS and SDSS^{93–95}, although at comparatively low statistical significance.

A recent study using Virgo's earlier Hubble volume simulations showed that the baryon wiggles should also be detectable in galaxy cluster samples⁹⁶. The length scale of the wiggles is a 'standard ruler' which, when observed at different redshifts, constrains the geometry and expansion history of the Universe and thus the dark energy equation of state. An example of what may be possible in the future is illustrated

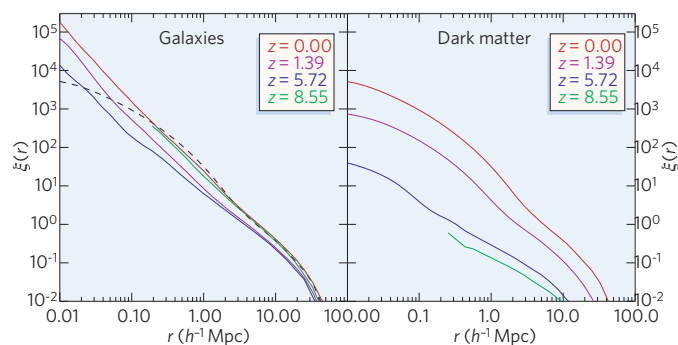


Figure 5 | Two-point correlation function of galaxies and dark matter at different epochs, in the Millennium simulation of structure formation⁵. The panel on the left gives the I-band galaxy correlation function ξ (selected according to $M_1 - 5 \log h < -20$ in the rest-frame) at redshifts $z = 8.55$, $z = 5.72$, $z = 1.39$ and $z = 0$ (corresponding to the epochs depicted in Fig. 4). The panel on the right shows the dark matter correlation functions at the same epochs. For comparison, the present-day dark matter correlation function is also drawn as a dashed line in the left panel. At $z = 8.55$, only data for $r > 200 h^{-1} \text{ kpc}$ are shown because the finite numerical resolution of the simulation precludes an accurate representation of the mass distribution on smaller scales than this at early times. The galaxy correlation function has a near power-law behaviour over several orders of magnitude and has almost equal strength at $z = 8.55$ and $z = 0$. By contrast, the dark matter correlation function grows by a large factor over this time span, and has a different shape from the galaxy correlation function.

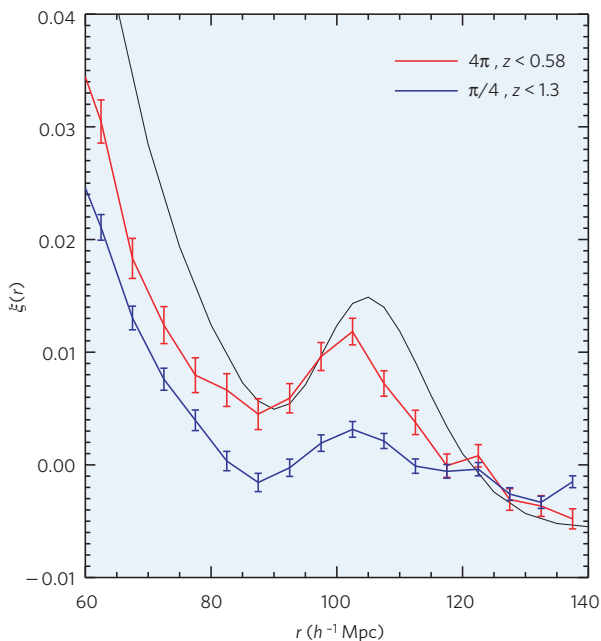


Figure 6 | The large-scale autocorrelation function of rich clusters. The two curves give the correlation function of clusters with X-ray temperature $kT > 5$ keV in light-cones constructed from the Hubble volume Λ CDM simulation¹⁰¹. The red line shows results for 124,000 clusters in a spherical light-cone out to $z = 0.58$, and the blue line shows results for 190,000 clusters in a light-cone of opening angle $\pi/2$ extending out to $z = 1.3$. The error bars are Poisson errors. The black line shows the results of linear theory scaled by the bias appropriate for the $z = 0.58$ sample. Nonlinear effects are responsible for the slight displacement of the position of the bump in the simulations relative to the position given by linear theory. Figure courtesy of Raul Angulo.

in Fig. 6, which shows the autocorrelation function of galaxy clusters in light-cones constructed from the Hubble volume Λ CDM simulation. The bump visible at a separation of $\sim 100 h^{-1}$ Mpc is the baryon feature that translates into a series of peaks when Fourier-transformed to give the power spectrum. New generations of galaxy and cluster surveys will target these oscillations and use them to constrain the evolution of dark energy.

In the more distant future, there are hopes that one day we will be able to probe the inflationary epoch directly by detecting the predicted background of gravitational waves^{97,98}. Not only would this provide strong evidence that inflation really happened but it would also rule out certain cosmological models inspired by string theory in which the collision of branes leads to the formation of our Universe. These predict a very weak gravitational wave background⁹⁹.

In the meantime, astrophysical studies of large-scale structure will continue to grow and to diversify, focusing on new issues such as the nature and evolution of nonlinear structure during the first billion years where we currently have no direct observations. No doubt new observations will continue to surprise us. Today, through the joint mysteries of dark matter and dark energy, cosmology arguably poses some of the most fundamental and exciting challenges of contemporary science. ■

- Colless, M. *et al.* The 2dF Galaxy Redshift Survey: spectra and redshifts. *Mon. Not. R. Astron. Soc.* **328**, 1039–1063 (2001).
- York, D. G. *et al.* The Sloan Digital Sky Survey: Technical summary. *Astron. J.* **120**, 1579–1587 (2000).
- Geller, M. J. & Huchra, J. P. Mapping the universe. *Science* **246**, 897–903 (1989).
- Bond, J. R., Kofman, L. & Pogosyan, D. How filaments of galaxies are woven into the cosmic web. *Nature* **380**, 603 (1996).
- Springel, V. *et al.* Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature* **435**, 629–636 (2005).
- White, S. D. M., Frenk, C. S. & Davis, M. Clustering in a neutrino-dominated universe. *Astrophys. J. Lett.* **274**, L1–L5 (1983).
- Guth, A. H. Inflationary universe: A possible solution to the horizon and flatness problems. *Physical Review D* **23**, 347–356 (1981).

- Starobinsky, A. A. Dynamics of phase transition in the new inflationary universe scenario and generation of perturbations. *Phys. Lett. B* **117**, 175–178 (1982).
- Zwicky, F. Die Rotverschiebung von extragalaktischen Nebeln. *Helv. Phys. Acta* **6**, 110–127 (1933).
- Zwicky, F. Nebulae as gravitational lenses. *Phys. Rev.* **51**, 290 (1937).
- Fischer, P. *et al.* Weak lensing with Sloan Digital Sky Survey commissioning data: the galaxy-mass correlation function to $1 h^{-1}$ Mpc. *Astron. J.* **120**, 1198–1208 (2000).
- Wilson, G., Kaiser, N., Luppino, G. A. & Cowie, L. L. Galaxy halo masses from galaxy-galaxy lensing. *Astrophys. J.* **555**, 572–584 (2001).
- Clowe, D., Luppino, G. A., Kaiser, N. & Gioia, I. M. Weak lensing by high-redshift clusters of galaxies. I. Cluster mass reconstruction. *Astrophys. J.* **539**, 540–560 (2000).
- Van Waerbeke, L. *et al.* Cosmic shear statistics and cosmology. *Astroparticle Phys.* **374**, 757–769 (2001).
- Kaiser, N. On the spatial correlations of Abell clusters. *Astrophys. J. Lett.* **284**, L9–L12 (1984).
- Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M. The evolution of large-scale structure in a universe dominated by cold dark matter. *Astrophys. J.* **292**, 371–394 (1985).
- Bardeen, J. M., Bond, J. R., Kaiser, N. & Szalay, A. S. The statistics of peaks of Gaussian random fields. *Astrophys. J.* **304**, 15–61 (1986).
- White, S. D. M., Navarro, J. F., Evrard, A. E. & Frenk, C. S. The baryon content of galaxy clusters — a challenge to cosmological orthodoxy. *Nature* **366**, 429 (1993).
- Allen, S. W., Schmidt, R. W., Fabian, A. C. & Ebeling, H. Cosmological constraints from the local X-ray luminosity function of the most X-ray-luminous galaxy clusters. *Mon. Not. R. Astron. Soc.* **342**, 287–298 (2003).
- Eke, V. R., Cole, S., Frenk, C. S. & Patrick Henry, J. Measuring Ω_0 using cluster evolution. *Mon. Not. R. Astron. Soc.* **298**, 1145–1158 (1998).
- Borgani, S. *et al.* Measuring Ω_m with the ROSAT Deep Cluster Survey. *Astrophys. J.* **561**, 13–21 (2001).
- Spergel, D. N. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: determination of cosmological parameters. *Astrophys. J. Suppl.* **148**, 175–194 (2003).
- Efstathiou, G., Sutherland, W. J. & Maddox, S. J. The cosmological constant and cold dark matter. *Nature* **348**, 705–707 (1990).
- Saunders, W., Frenk, C., Rowan-Robinson, M., Lawrence, A. & Efstathiou, G. The density field of the local universe. *Nature* **349**, 32–38 (1991).
- Perlmutter, S. *et al.* Measurements of omega and lambda from 42 high-redshift supernovae. *Astrophys. J.* **517**, 565–586 (1999).
- Riess, A. G. *et al.* Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116**, 1009–1038 (1998).
- Smoot, G. F. *et al.* Structure in the COBE differential microwave radiometer first-year maps. *Astrophys. J. Lett.* **396**, L1–L5 (1992).
- de Bernardis, P. *et al.* A flat Universe from high-resolution maps of the cosmic microwave background radiation. *Nature* **404**, 955–959 (2000).
- Hanany, S. *et al.* MAXIMA-1: A measurement of the cosmic microwave background anisotropy on angular scales of 10° . *Astrophys. J. Lett.* **545**, L5–L9 (2000).
- Netterfield, C. B. *et al.* A measurement by BOOMERANG of multiple peaks in the angular power spectrum of the cosmic microwave background. *Astrophys. J.* **571**, 604–614 (2002).
- Kovac, J. M. *et al.* Detection of polarization in the cosmic microwave background using DASI. *Nature* **420**, 772–787 (2002).
- Leitch, E. M. *et al.* Measurement of polarization with the Degree Angular Scale Interferometer. *Nature* **420**, 763–771 (2002).
- Contaldi, C. R., Hoekstra, H. & Lewis, A. Joint cosmic microwave background and weak lensing analysis: constraints on cosmological parameters. *Phys. Rev. Lett.* **90**, 221303 (2003).
- Tegmark, M. *et al.* The three-dimensional power spectrum of galaxies from the Sloan Digital Sky Survey. *Astrophys. J.* **606**, 702–740 (2004).
- Sánchez, A. G. *et al.* Cosmological parameters from cosmic microwave background measurements and the final 2dF Galaxy Redshift Survey power spectrum. *Mon. Not. R. Astron. Soc.* **366**, 189–207 (2006).
- Seljak, U. *et al.* Cosmological parameter analysis including SDSS Ly- α forest and galaxy bias: Constraints on the primordial spectrum of fluctuations, neutrino mass, and dark energy. *Phys. Rev. D* **71**, 103515 (2005).
- Sunyaev, R. A. & Zeldovich, Y. B. Small-scale fluctuations of relic radiation. *Astrophys. Space Sci.* **7**, 3–19 (1970).
- Cen, R., Miralda-Escudé, J., Ostriker, J. P. & Rauch, M. Gravitational collapse of small-scale structure as the origin of the Lyman- α forest. *Astrophys. J. Lett.* **437**, L9–L12 (1994).
- White, S. D. M. & Frenk, C. S. Galaxy formation through hierarchical clustering. *Astrophys. J.* **379**, 52–79 (1991).
- Press, W. H. & Schechter, P. Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. *Astrophys. J.* **187**, 425–438 (1974).
- Lacey, C. & Cole, S. Merger rates in hierarchical models of galaxy formation. *Mon. Not. R. Astron. Soc.* **262**, 627–649 (1993).
- White, S. D. M. & Rees, M. J. Core condensation in heavy halos: A two-stage theory for galaxy formation and clustering. *Mon. Not. R. Astron. Soc.* **183**, 341–358 (1978).
- Reed, D. S. *et al.* The first generation of star-forming haloes. *Mon. Not. R. Astron. Soc.* **363**, 393–404 (2005).
- Ciardi, B., Ferrara, A. & White, S. D. M. Early reionization by the first galaxies. *Mon. Not. R. Astron. Soc.* **344**, L7–L11 (2003).
- Abel, T., Bryan, G. L. & Norman, M. L. The formation of the first star in the Universe. *Science* **295**, 93–98 (2002).
- Hernquist, L., Katz, N., Weinberg, D. H. & Miralda-Escudé, J. The Lyman- α forest in the cold dark matter model. *Astrophys. J. Lett.* **457**, L51 (1996).
- Croft, R. A. C., Weinberg, D. H., Katz, N. & Hernquist, L. Recovery of the power spectrum of mass fluctuations from observations of the Ly- α Forest. *Astrophys. J.* **495**, 44 (1998).
- Croft, R. A. C. *et al.* Toward a precise measurement of matter clustering: Ly- α forest data at redshifts 2–4. *Astrophys. J.* **581**, 20–52 (2002).
- Kim, T.-S., Viel, M., Haehnelt, M. G., Carswell, R. F. & Cristiani, S. The power spectrum

- of the flux distribution in the Lyman α forest of a large sample of UVES QSO absorption spectra (LUQAS). *Mon. Not. R. Astron. Soc.* **347**, 355–366 (2004).
50. Viel, M., Haehnelt, M. G. & Springel, V. Inferring the dark matter power spectrum from the Lyman α forest in high-resolution QSO absorption spectra. *Mon. Not. R. Astron. Soc.* **354**, 684–694 (2004).
 51. McDonald, P., Seljak, U., Cen, R., Bode, P. & Ostriker, J. P. Physical effects on the Ly- α forest flux power spectrum: damping wings, ionizing radiation fluctuations and galactic winds. *Mon. Not. R. Astron. Soc.* **360**, 1471–1482 (2005).
 52. Aguirre, A. *et al.* Confronting Cosmological Simulations with Observations of Intergalactic Metals. *Astrophys. J. Lett.* **620**, L13–L17 (2005).
 53. Kaiser, N. Weak gravitational lensing of distant galaxies. *Astrophys. J.* **388**, 272–286 (1992).
 54. Van Waerbeke, L., Mellier, Y. & Hoekstra, H. Dealing with systematics in cosmic shear studies: New results from the VIRMOS–Descart survey. *Astrophys. J.* **629**, 75–84 (2005).
 55. Mandelbaum, R., Seljak, U., Kauffmann, G., Hirata, C. M. & Brinkmann, J. Galaxy halo masses and satellite fractions from galaxy–galaxy lensing in the SDSS: stellar mass, luminosity, morphology, and environment dependencies. *ArXiv Astrophys. e-prints* <arXiv:astro-ph/0511164> (2005).
 56. Viel, M., Weller, J. & Haehnelt, M. G. Constraints on the primordial power spectrum from high-resolution Lyman α forest spectra and WMAP. *Mon. Not. R. Astron. Soc.* **355**, L23–L28 (2004).
 57. Benson, A. J., Cole, S., Frenk, C. S., Baugh, C. M. & Lacey, C. G. The nature of galaxy bias and clustering. *Mon. Not. R. Astron. Soc.* **311**, 793–808 (2000).
 58. Masjedi, M. *et al.* Very small-scale clustering and merger rate of luminous red galaxies. *ArXiv Astrophys. e-prints* <arXiv:astro-ph/0512166> (2005).
 59. Croton, D. J. *et al.* The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. *Mon. Not. R. Astron. Soc.* **365**, 11–28 (2006).
 60. De Lucia, G., Springel, V., White, S. D. M., Croton, D. & Kauffmann, G. The formation history of elliptical galaxies. *Mon. Not. R. Astron. Soc.* **366**, 499–509 (2006).
 61. Giavalisco, M. *et al.* The angular clustering of Lyman-break galaxies at redshift $z \sim 3$. *Astrophys. J.* **503**, 543 (1998).
 62. Adelberger, K. L. *et al.* A counts-in-cells analysis of Lyman-break galaxies at redshift $z \sim 3$. *Astrophys. J.* **505**, 18–24 (1998).
 63. Mo, H. J. & Fukugita, M. Constraints on the cosmic structure formation models from early formation of giant galaxies. *Astrophys. J. Lett.* **467**, L9 (1996).
 64. Baugh, C. M., Cole, S., Frenk, C. S. & Lacey, C. G. The epoch of galaxy formation. *Astrophys. J.* **498**, 504 (1998).
 65. Weinberg, S. The cosmological constant problem. *Rev. Mod. Phys.* **61**, 1–23 (1989).
 66. Hoeflich, P., Wheeler, J. C. & Thielemann, F. K. Type Ia supernovae: influence of the initial composition on the nucleosynthesis, light curves, and spectra and consequences for the determination of Ω_m and Λ . *Astrophys. J.* **495**, 617 (1998).
 67. Travaglio, C., Hillebrandt, W. & Reinecke, M. Metallicity effect in multi-dimensional SN Ia nucleosynthesis. *Astroparticle Phys.* **443**, 1007–1011 (2005).
 68. Hamuy, M. *et al.* The absolute luminosities of the Calan/Tololo Type Ia supernovae. *Astron. J.* **112**, 2391 (1996).
 69. Gallagher, J. S. *et al.* Chemistry and star formation in the host galaxies of Type Ia supernovae. *Astrophys. J.* **634**, 210–226 (2005).
 70. Blanchard, A., Douspis, M., Rowan-Robinson, M. & Sarkar, S. An alternative to the cosmological ‘concordance model’. *Astrophys. J.* **412**, 35–44 (2003).
 71. Chiang, L.-Y., Naselsky, P. D., Verkhodanov, O. V. & Way, M. J. Non-gaussianity of the derived maps from the First-year Wilkinson Microwave Anisotropy Probe data. *Astrophys. J. Lett.* **590**, L65–L68 (2003).
 72. Vielva, P., Martínez-González, E., Barreiro, R. B., Sanz, J. L. & Cayón, L. Detection of non-gaussianity in the Wilkinson Microwave Anisotropy Probe First-Year data using spherical wavelets. *Astrophys. J.* **609**, 22–34 (2004).
 73. de Oliveira-Costa, A., Tegmark, M., Zaldarriaga, M. & Hamilton, A. Significance of the largest scale CMB fluctuations in WMAP. *Phys. Rev. D* **69**, 063516 (2004).
 74. Eriksen, H. K., Hansen, F. K., Banday, A. J., Górski, K. M. & Lilje, P. B. Asymmetries in the cosmic microwave background anisotropy field. *Astrophys. J.* **605**, 14–20 (2004).
 75. Land, K. & Magueijo, J. Examination of evidence for a preferred axis in the cosmic radiation anisotropy. *Phys. Rev. Lett.* **95**, 071301 (2005).
 76. Jaffe, T. R., Banday, A. J., Eriksen, H. K., Górski, K. M. & Hansen, F. K. Evidence of vorticity and shear at large angular scales in the WMAP data: a violation of cosmological isotropy? *Astrophys. J. Lett.* **629**, L1–L4 (2005).
 77. Aharonian, F. *et al.* Very high energy gamma rays from the direction of Sagittarius A*. *Astrophys. J.* **425**, L13–L17 (2004).
 78. Bergström, L., Ullio, P. & Buckley, J. H. Observability of gamma rays from dark matter neutralino annihilations in the Milky Way halo. *Astrophys. J.* **9**, 137–162 (1998).
 79. Bekenstein, J. D. Relativistic gravitation theory for the modified newtonian dynamics paradigm. *Phys. Rev. D* **70**, 083509 (2004).
 80. Aguirre, A., Schaye, J. & Quataert, E. Problems for modified newtonian dynamics in clusters and the Ly- α forest? *Astrophys. J.* **561**, 550–558 (2001).
 81. Klypin, A., Kravtsov, A. V., Valenzuela, O. & Prada, F. Where are the missing galactic satellites? *Astrophys. J.* **522**, 82–92 (1999).
 82. Moore, B. *et al.* Dark matter substructure within galactic halos. *Astrophys. J. Lett.* **524**, L19–L22 (1999).
 83. Navarro, J. F., Frenk, C. S. & White, S. D. M. A Universal density profile from hierarchical clustering. *Astrophys. J.* **490**, 493 (1997).
 84. de Blok, W. J. G., McGaugh, S. S., Bosma, A. & Rubin, V. C. Mass density profiles of low surface brightness galaxies. *Astrophys. J. Lett.* **552**, L23–L26 (2001).
 85. Hayashi, E. *et al.* The inner structure of Λ CDM haloes. II. Halo mass profiles and low surface brightness galaxy rotation curves. *Mon. Not. R. Astron. Soc.* **355**, 794–812 (2004).
 86. Bullock, J. S., Kravtsov, A. V. & Weinberg, D. H. Hierarchical galaxy formation and substructure in the Galaxy’s stellar halo. *Astrophys. J.* **548**, 33–46 (2001).
 87. Benson, A. J., Frenk, C. S., Lacey, C. G., Baugh, C. M. & Cole, S. The effects of photoionization on galaxy formation. II. Satellite galaxies in the Local Group. *Mon. Not. R. Astron. Soc.* **333**, 177–190 (2002).
 88. Kochanek, C. S. & Dalal, N. Tests for substructure in gravitational lenses. *Astrophys. J.* **610**, 69–79 (2004).
 89. Oguri, M., Takada, M., Umetsu, K. & Broadhurst, T. Can the steep mass profile of A1689 be explained by a triaxial dark halo? *Astrophys. J.* **632**, 841–846 (2005).
 90. Pierce, A. Dark matter in the finely tuned minimal supersymmetric standard model. *Phys. Rev. D* **70**, 075006 (2004).
 91. Haiman, Z., Mohr, J. J. & Holder, G. P. Constraints on cosmological parameters from future galaxy cluster surveys. *Astrophys. J.* **553**, 545–561 (2001).
 92. Peebles, P. J. E. & Yu, J. T. Primeval adiabatic perturbation in an expanding universe. *Astrophys. J.* **162**, 815 (1970).
 93. Cole, S. *et al.* The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications. *Mon. Not. R. Astron. Soc.* **362**, 505–534 (2005).
 94. Eisenstein, D. J. *et al.* Detection of the baryon acoustic peak in the large-scale correlation function of SDSS luminous red galaxies. *Astrophys. J.* **633**, 560–574 (2005).
 95. Huetsi, G. Acoustic oscillations in the SDSS DR4 Luminous Red Galaxy sample power spectrum. *ArXiv Astrophys. e-prints* <arXiv:astro-ph/0512201> (2005).
 96. Angulo, R. *et al.* Constraints on the dark energy equation of state from the imprint of baryons on the power spectrum of clusters. *Mon. Not. R. Astron. Soc.* **362**, L25–L29 (2005).
 97. Allen, B. Stochastic gravity-wave background in inflationary–universe models. *Phys. Rev. D* **37**, 2078–2085 (1988).
 98. Lyth, D. H. What would we learn by detecting a gravitational wave signal in the cosmic microwave background anisotropy? *Phys. Rev. Lett.* **78**, 1861–1863 (1997).
 99. Boyle, L. A., Steinhardt, P. J. & Turok, N. Cosmic gravitational-wave background in a cyclic universe. *Phys. Rev. D* **69**, 127302 (2004).
 100. Gott, J. R. I. *et al.* A map of the Universe. *Astrophys. J.* **624**, 463–484 (2005).
 101. Evrard, A. E. *et al.* Galaxy clusters in Hubble volume simulations: cosmological constraints from sky survey populations. *Astrophys. J.* **573**, 7–36 (2002).

Acknowledgements We thank L. van Waerbeke for providing the data of Fig. 3, and R. Angulo for preparing Fig. 6.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to V.S. (vspringel@mpa-garching.mpg.de).

High-redshift galaxy populations

Esther M. Hu¹ & Lennox L. Cowie¹

We now see many galaxies as they were only 800 million years after the Big Bang, and that limit may soon be exceeded when wide-field infrared detectors are widely available. Multi-wavelength studies show that there was relatively little star formation at very early times and that star formation was at its maximum at about half the age of the Universe. A small number of high-redshift objects have been found by targeting X-ray and radio sources and most recently, γ -ray bursts. The γ -ray burst sources may provide a way to reach even higher-redshift galaxies in the future, and to probe the first generation of stars.

Over the past decade, the availability of a new generation of ground- and space-based instruments has transformed our understanding of early galaxies. Before this, observational studies of the star-forming components in the early Universe were limited by both the extreme faintness of these distant objects and the sparseness of their distribution on the sky. At these early times in the life of the Universe, a substantial fraction of the matter is in neutral hydrogen that has not yet been assembled into galaxies. The effect of a largely neutral intergalactic medium (IGM), one that has not been substantially ionized by star-forming galaxies or quasars, is to further attenuate the light from this early luminous population. It does this through absorption and resonant scattering by intervening hydrogen gas along the line of sight — the famed Gunn–Peterson effect¹.

The pace of recent progress has been stunning. Before 1996 no normal star-forming galaxy was known to have a redshift remotely approaching that of the most distant quasars and radio galaxies, but now not only is it possible to confirm redshifts of distant galaxies that correspond to cosmic times within the first billion years of the age of the Universe, but such objects outdo quasars in both redshift record and known number. Large telescopes such as the 10-m Keck telescopes, the 8-m Very Large Telescopes (VLTs) of the European Southern Observatory, the 8.3-m Subaru telescope and the 8-m Gemini North and Gemini South telescopes make redshift determinations possible for such early galaxies, which were too faint to be spectroscopically identified by earlier generations of telescopes. The redshift limit for studies of star-forming galaxies jumped to $z = 4.55$ and $z = 4.69$ (refs 2, 3), or a cosmic age of about 1.3 Gyr, in 1996, soon after the first Keck 10-m telescope went into operation. These values may be compared to the quasar redshift record⁴ for that time of $z = 4.89$, which was set in 1991. Over the next two years, the galaxy redshift limit quickly passed^{5–7} beyond $z = 5$ to redshift $z = 5.74$, or about 970 million years after the Big Bang. The galaxy limit continued to exceed the redshifts of $z > 5$ quasars just then being found by the Sloan Digital Sky Survey until 2001 when the first quasars beyond $z = 5.8$ were reported⁸. However, the discovery⁹ of a galaxy of redshift $z = 6.56$ the following year once more surpassed the most distant quasar redshifts, and samples of galaxies in this redshift range have continued to grow rapidly^{10–15}. At present, the highest quasar redshift¹⁶ in the literature is $z = 6.42$, and the highest spectroscopically confirmed redshift¹³ for a galaxy is $z = 6.597$, which corresponds to a cosmic time about 800 million years after the Big Bang. Higher redshifts have been claimed^{17,18} but still await spectroscopic confirmation.

Two recent scientific developments have further spurred interest in the number of galaxies and quasars at early epochs and their contribution to the ionizing radiation. These are the studies^{19–22} of the IGM

towards the newly discovered $z > 5$ quasars found by the Sloan Digital Sky Survey, and the recent determination^{23,24}, made by the Wilkinson Microwave Anisotropy Probe (WMAP) space experiment, of a high optical depth due to electron scattering in the microwave background radiation. The latter result suggests that a substantial population of star-forming galaxies and quasars is already in place at very early times. Both measurements probe the state of the gaseous material that comprises most of the baryonic material at early times, before substantial galaxy formation has taken place. In the standard picture, the Universe cooled and expanded following the Big Bang, with ions recombining into neutral atoms at $z \approx 1,100$, about 370,000 years after the Big Bang²⁵. This material remained largely neutral until energy released from the formation of collapsed objects ionized the surrounding medium, rendering it transparent to ultraviolet light and allowing the first galaxies to become visible. This epoch of re-ionization may lie at $z \approx 6.2$, as indicated by quasar absorption-line studies^{19,20,22} that measure the increase in absorption troughs and opacity of the IGM as a function of redshift along the lines of sight to the distant quasars. But this result is still controversial.

However, the WMAP measurements of optical depths from the microwave background measurements suggested an earlier redshift of re-ionization at $z \approx 17 \pm 5$, or only a few hundred million years after the Big Bang. This has prompted the development of complex models^{25–27} for the ionization history of the Universe and the physical properties of the IGM, as well as consideration of spatial and structural variations^{28,29} that allow galaxies beyond $z \approx 6.2$ to be observed. In addition, it has focused attention on the sources of ionization, and on the galaxies and quasars that may be present at very early times. Current interest is directed at learning how many galaxies there may be at substantially higher redshifts, establishing their properties, and finding ways to extend the searches beyond the present redshift samples, perhaps through observations at other wavelengths.

An excellent review of high-redshift galaxy studies up to 2003 was written by Spinrad^{30,31}. Since then we have seen a rapid growth in the numbers of galaxies that populate the high-redshift ($z > 5$) end of the samples. This redshift range corresponds to roughly the first 1.2 Gyr after the Big Bang, and is the focus of the present review. The rapid growth has come from two developments: the colour analysis of the very deep Hubble Space Telescope (HST) images of the two Great Observatories Origins Deep Survey³² (GOODS) fields and the Hubble Ultra Deep Field (HUDF) subregion of the GOODS-South field; and the searches for galaxies with hydrogen Ly α emission from star formation that have been made possible by the giant charge-coupled device (CCD) cameras on large telescopes. A third technique that may be very powerful,

¹Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA.

locating the galaxies through γ -ray bursts, is still at a very preliminary stage. A few objects have been found at redshifts $z > 5$ from radio surveys³³ and X-ray surveys^{34,35}, all with redshifts below $z = 5.4$. Such high-redshift active galactic nuclei are estimated to make only a small contribution³⁴ to the ionizing background at these redshifts. Only two objects, the highest-redshift³³ radio galaxy TN J0924–2201 at $z = 5.2$ and the highest-redshift¹⁶ quasar SDSS J114816.64+525150.3 at $z = 6.42$ have so far been detected in molecular lines^{36,37}, as summarized in a recent review³⁸ of molecular gas observed at high redshift. However, we expect the advent of the Atacama Large Millimeter Array (ALMA) with its increased sensitivity to revolutionize such studies³⁸.

Despite the faintness of the optical galaxies and the attenuation of their light by the intervening IGM, these star-forming galaxies dominate the high-redshift population. Paradoxically, it is the very prevalence of neutral hydrogen in the high- z IGM that provides the key observational signature for such distant galaxy populations. Intervening neutral hydrogen gas scatters rest-frame ultraviolet continuum light at wavelengths below the break in the Lyman series, corresponding to a continuum

colour break below the Lyman α emission line, and an asymmetric line profile forms because of scattering of the blue wing of the Lyman α emission line^{28,39–41}. Both these signatures — the strong hydrogen Lyman α emission line excited by star formation and its asymmetric profile, and the sharp decrease in galaxy light transmitted by the IGM below the wavelength of the Lyman continuum series, also known as the Lyman continuum colour break — can be used to identify high-redshift galaxy candidates.

The large, spectroscopically complete samples now available from wide-area surveys for a number of discrete fields^{13,15,42,43} provide a means of studying early galaxies and star formation. These give a handle on the problem of cosmic variance, and also allow us to build up the large samples of data required for studies of the luminosity function. The spectroscopic information makes it possible to estimate the star-forming population at high redshift and its evolution. Detailed examination and comparison of the strength and profile shape of the Ly α emission line may provide insights into possible changes in the underlying stellar population or in the neutral hydrogen content of the surrounding intergalactic medium.

The immediate extension to current surveys will come from pushing present methods to increasingly red wavelengths and higher redshifts through the use of large-format infrared detector arrays. Higher-redshift studies may also detect the early, luminous, metal-free generation of Population III stars^{44,45}. Different techniques, such as the very promising method⁴⁶ of using γ -ray burst sources to identify⁴⁷ high-redshift galaxies, may also benefit from the increased visibility of luminous Population III star components⁴⁵.

Colour-selected surveys

Deep surveys carried out with the advanced camera for surveys (ACS) and the near-infrared camera NICMOS on the Hubble Space Telescope have enabled very high- z galaxies to be selected using the colour dropout technique, which uses the strength of the Lyman continuum break to select candidates according to the suppression of continuum light below a characteristic wavelength. At these redshifts the intergalactic gas is extremely optically thick to Lyman α scattering^{21,22}, and galaxies near a redshift of $z = 6$ can be fairly robustly found by looking for objects with deep breaks between the z' ($\sim 9,100$ Å) and i ($\sim 8,100$ Å) bands. Bunker and collaborators^{48–50} and Bouwens and collaborators^{18,51–53} have used this technique to determine the luminosity, functions and star-formation rates at these redshifts (see also Yan and Windhorst's analysis of the HUDF⁵⁴). These objects are faint, with equivalent visual magnitudes of $AB \approx 25$ or more, but some confirming spectroscopy has been obtained^{55–57}. Low-resolution spectroscopy using the grism (a combination of diffraction grating and prism) on the ACS camera has confirmed the break in a larger number of objects in the HUDF⁵⁸, and narrowband imaging of the GOODS-North region has found a number of Ly α emitters⁴³.

Bouwens and collaborators⁵³ have extended this into the near-infrared using NICMOS observations of the HUDF. There is essentially no confirmation of these objects, so the results are based solely on the robustness of the colour dropout technique. They may best be viewed as upper limits, at least at the very highest redshifts ($z \approx 10$). For bright sources at $z \approx 5–6$ in the HUDF, Yan and collaborators⁵⁴ have extended analyses with Spitzer's Infrared Array Camera (IRAC) imaging.

We have summarized these results in Fig. 1. Rather than plotting these as $d\rho/dt$ versus redshift as is usually done, we plot the quantity $d\rho/dt \times t$ versus the cosmic time t since the Big Bang. Here $d\rho/dt$ is the star-formation rate per co-moving volume computed using a Salpeter initial mass function extending to $0.1 M_{\odot}$. The advantage of this type of display is that we can see more directly how many stars are formed at a given time. The blue points on the figure show the direct star-formation measurement from the ultraviolet light without any correction for extinction, and this is compared to similar results at lower redshift. After some early debate on the high-redshift value there now seems a fair agreement in the determinations. At low redshift, too, most of the measurements agree well. (The one exception is the local determination, where the new

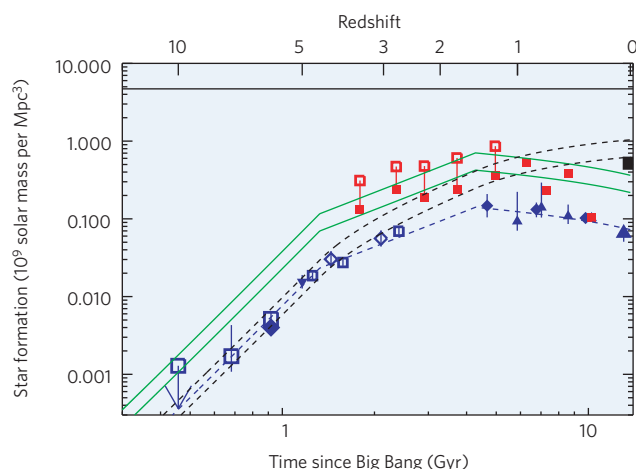


Figure 1 | Star formation history versus cosmic time since the Big Bang.

Cosmic time is computed for a flat geometry with $\Omega_M = 0.3$ and $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. We plot the star formation rate computed for a Salpeter initial mass function (IMF) extending to $0.1 M_{\odot}$ multiplied by the cosmic time, as this shows more clearly what fraction of the baryonic mass in stars is created at any time. The data combine multi-wavelength data from far-ultraviolet to submillimetre and radio wavelengths, to sample the star formation rate at a variety of redshifts. The blue symbols show the star formation, which is directly seen at rest-frame ultraviolet wavelengths. At each redshift we have tried to choose some of the best determinations of this quantity rather than all available results. At late times we show the local determination from the Galaxy Evolution Explorer (GALEX) (large solid triangle) by Wyder *et al.*⁵⁹, which is lower than previous measurements, the GALEX determinations of Schiminovich *et al.*⁹⁵ (small solid triangles) and the ground-based determinations of Wilson *et al.*⁹⁶ (small solid diamonds) which are the most accurate measurements near $z = 1$. At intermediate times we show the values obtained by Steidel *et al.*⁸² (open diamonds), Bouwens *et al.*⁵¹ (small open squares), and Iwata *et al.*⁹⁷ (solid inverted triangle). At the earliest times we show the results of Bunker *et al.*⁵⁰ (large solid diamond) and Bouwens *et al.*⁵³ (large open squares); these results are now in much closer agreement than previous determinations by these groups. The dashed line shows the parameterization given in the text. The solid green lines show the total star formation rates that would be inferred if we apply reasonable extinction corrections to the ultraviolet light (upper curve = 5, lower curve = 3). The solid red symbols show the recent direct measurements of the total star formation rate based on submillimetre and radio observations from Wang *et al.*⁸⁷; the open red symbols show the maximal correction for incompleteness in the range $z = 1–3$ for this data. There is broad agreement now with the dust-corrected ultraviolet determination, although the populations giving rise to the light in the far infrared are somewhat disjoint from those giving rise to the ultraviolet light. The solid black line shows the cosmic baryon density, the black dashed lines the cumulative star formation and the solid black square the present-day baryon density in galaxies as estimated by Cole *et al.*⁹⁸.

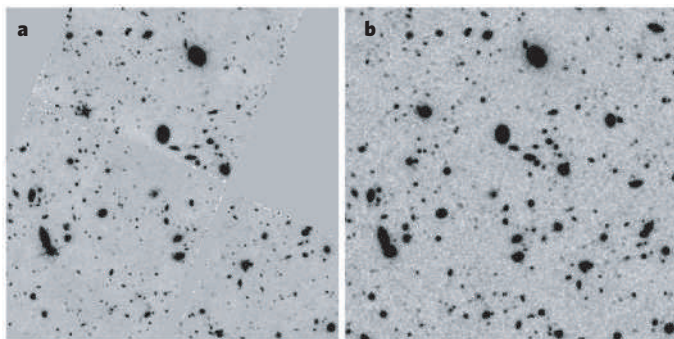


Figure 2 | The Hubble Deep Field used for HST studies and for ground-based narrowband Ly α surveys. **a**, The Hubble Deep Field imaged with the Wide-Field Planetary Camera 2 (WFPC2) on HST through the F814W continuum filter. **b**, The same region imaged from the ground through a narrowband NB816 filter of width 120 Å (for Ly α emission at $z \approx 5.7$) using the Suprime-Cam⁶⁷ mosaic CCD camera on Subaru. The narrowband images that can be obtained with this camera are nearly comparable in depth to the Hubble Deep Field, but cover more than a hundred times the area.

measurement using the GALEX satellite⁵⁹ is lower than the previous measurement made using the FOCA balloon-borne data⁶⁰.)

We can make a simple segmented power-law fit to the evolution of star formation, which is well parameterized by the form $dp/dt = 0.0053(1+z)^2$ (for $z < 1.5$), $dp/dt = 0.068(1+z)^{-0.5}$ (for $1.5 < z < 4.5$) and $dp/dt = 37.7(1+z)^4$ (for $z > 4.5$) $M_{\odot} \text{ Mpc}^{-3} \text{ yr}^{-1}$. The standard correction of a multiplicative factor of about 3–5 for dust extinction produces good agreement with measurements from radio and submillimetre observations at intermediate redshifts. Applying it uniformly we can map the total star production versus time, which is shown by the dashed black lines. The lower estimate for the dust-corrected star formation maps directly onto the present galaxy mass density, whereas the higher estimate slightly overpredicts this quantity and might require an initial mass function that is more weighted to high-mass stars at some periods of the cosmic time.

It is striking how little of the star formation has occurred at early times, according to these star-formation rates from these colour-selected samples. Less than 0.007 of the total star formation has taken place before $z = 6$ and, even with the high-end dust correction, less than 0.0014 of the baryons have been incorporated into stars. This raises some fundamental problems such as whether there are enough photons available to ionize the intergalactic gas⁵⁰. A simple way of seeing this problem is to note that even if all the nucleosynthetic energy ($0.007mc^2$ per stellar baryon) were instantly converted to 13.6 eV photons, we would produce only 750 ionizing photons per baryon and at most a few per cent of these will escape the galaxy. Thus, although this particular problem can be solved within the uncertainties in the initial mass function, the metallicity of the stars and the escape fraction of ionizing photons from the galaxies⁶¹, it is difficult enough that we should ask whether we are undercounting the amount of star formation with this technique.

Cosmic variance is a possible problem^{62,63} because the fields are small. This will generally result in an underestimate because the bulk of the sky area is under-dense, and it is only when we include the smaller number of over-dense regions that we average properly. Alternatively, much of the star formation could be in low-luminosity objects lying below the sensitivity of the observations^{26,27}, or we may be underestimating the magnitudes of objects or missing some objects because of the effects of surface brightness dimming⁵¹. It is therefore important to look at other approaches to the problem.

Lyman α surveys

The most successful method of finding galaxies that can be spectroscopically identified at redshifts beyond $z \approx 5$, where galaxy continua are faint against a strong night-sky background, is by using the redshifted hydrogen Ly α emission line. The imaging form of the discovery technique typically uses narrowband filters with central wavelengths

positioned at locally dark regions of the night sky free of strong atmospheric airglow lines. The filter samples the redshifted Ly α emission feature displaced by a factor $z \times \lambda_0$ from the rest-frame wavelength of the line ($\lambda_0 = 1,216 \text{ Å}$), where z is the redshift of the detected high-redshift galaxies. Lyman α is present in emission only during the early stages of a star-formation episode ($< 5 \times 10^8$ years⁶⁴) and, because of its complex radiative transfer, is easily destroyed by dust. Nevertheless, the technique successfully picks up many objects, and at lower redshifts ($z \approx 3$) finds about 20% of the colour-selected galaxies⁶⁵. Because the background night sky contains many strong airglow emission lines at red wavelengths, narrowband surveys in preselected locally dark wavelength regions are also a good way to study the statistics of these distant galaxies and to compare how their number density and luminosity evolves with redshift. The great advantages of the Ly α technique are that the necessary sensitivity can be obtained with deep ground-based observations, either with narrowband filters or spectroscopic observations, and that very large areas of sky may be searched.

The current generation of wide-field Ly α surveys was driven by the development of wide-field, red-sensitive instruments on the large telescopes, first used on the 4-m class telescopes⁶⁶, and then on the Suprime-Cam mosaic CCD imager⁶⁷, which has a half-degree field of view, on the 8.3-m Subaru Telescope. Long exposures on this instrument can reach sensitivities in narrowband observations that are quite close to the space-based imaging observations, as we illustrate in Fig. 2. The narrowband imaging is used to select line emitters, and deep multicolour data are then used to select out those objects that are high-redshift galaxies with strong Ly α emission and continuum colour breaks. Most of these investigations use narrow bandpass ($\sim 100 \text{ Å}$) filters centred at 8,150 Å and 9,130 Å (Ly α at $z \approx 5.7$ and $z \approx 6.5$), which take advantage of the gaps between OH bands in the night sky where the sky background is very low.

The Lyman α spectral lines have a very distinctive shape because the blue side of the line is removed by the scattering in the IGM, and this red asymmetry makes them easy to recognize, as is illustrated in Fig. 3. More than 100 emitters have now been identified for the redshift range^{6,7,15,42,43,55,56,68–76} $z = 5.7–6.3$ and some 40 to 50 for redshifts^{9–15} $z \approx 6.6$. Surprisingly, there is very little change in the shape and equivalent width of the emitters or in their luminosity function with redshifts from $z = 3$ out to $z = 6.5$, suggesting that the selected objects continue to lie in predominantly ionized regions and that much of the volume of the IGM is still ionized at these times. Malhotra *et al.*⁷⁷ have combined this

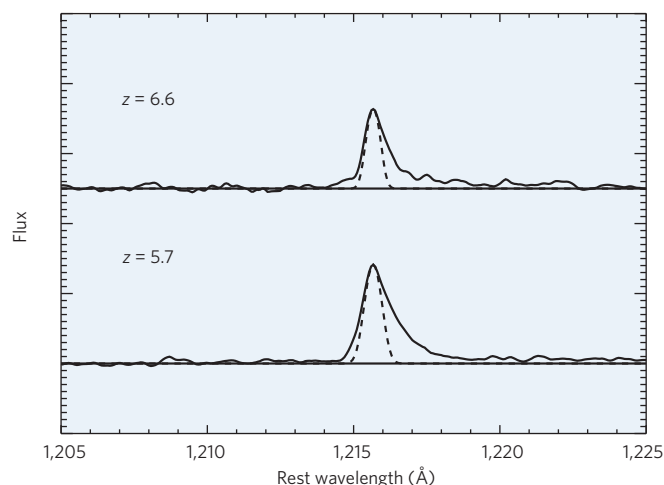


Figure 3 | Lyman α emission-line profiles of galaxies at $z \approx 5.7$ and $z \approx 6.6$. These are stacked profiles of galaxies identified at these redshifts¹⁵. The dotted profile shows the spectral resolution from the night sky lines. The steep blue fall-off due to absorption by neutral hydrogen is clearly evident in the asymmetric line profile. The profiles are extremely similar in equivalent width (56 Å and 60 Å) and shape at both redshifts. Both the line properties and the Ly α luminosity function appear to show little evolution from $z = 3$ to $z = 6.6$ (refs 9,15,42,68,90,99,100).

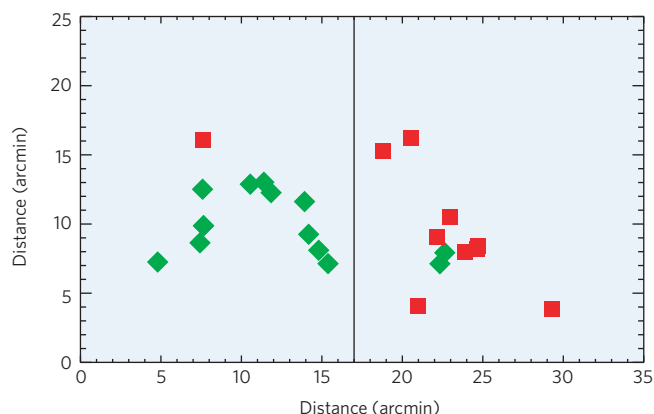


Figure 4 | Distribution of $z \approx 5.7$ and $z \approx 6.5$ in two adjacent Suprime-Cam fields. The distribution of galaxies is highly structured. Here we show the distribution of galaxies in the redshift range $z = 5.6$ – 5.8 (green diamonds) and $z = 6.5$ – 6.7 (red squares) in an area of just under 900 square arcminutes covered by two Suprime-Cam images (shown by the solid vertical line). The galaxies appear to lie in extended filaments, and even on fields as large as these there is considerable field-to-field variance^{15,42,101}. Here one field has a large number of galaxies at redshift 6.5 and few at redshift 5.7, whereas in the neighbouring field this is reversed.

requirement that the individual objects must lie in large ionized regions with the number density of Lyman α emitters at $z = 6.5$ to show that a minimum of 20% of the IGM at these redshifts must still be substantially ionized. Wyithe and Loeb^{26,27} and Furlanetto *et al.*²⁹ argue that clustering effects from low-luminosity emitters and the structured distribution of these sources may weaken such arguments, but the distribution of emitters is sparse enough (Fig. 4) that the arguments cannot substantially underestimate the filling factors of ionized regions.

At continuum magnitudes brighter than $AB \approx 25.5$, the number density of Lyman-emitting galaxies at $z \approx 6$ is roughly 1,300 (refs 42, 43) per square degree per unit redshift, or of the order of 100 objects per square degree for typical filter bandpasses (see also Stern *et al.*¹⁴ for a summary of other recent estimates) and the distribution is highly correlated on sub-degree scales, often lying in apparently filamentary structures tracing the underlying cosmic web (Fig. 4). There are very large amounts of variance on the scale of the size of the Suprime field, roughly 600 square arcminutes. This emphasizes the need to cover very large areas if we are to measure the star-formation rates properly at these times.

The problem of cosmic variance is also an issue for the success of gravitational lensing searches^{9,17,71,78,79} for high-redshift objects. These potentially provide detailed information on lower luminosity high-redshift galaxies, but over a very restricted sampling region. Lensing studies have successfully identified high-redshift galaxies^{9,17} approaching $z \approx 7$, but identification of a $z \approx 10$ galaxy⁷⁹ is controversial^{80,81}.

With the high-redshift samples in hand, we can construct luminosity functions at $z \approx 5.7$ and $z \approx 6.5$, and compare these with results at lower redshifts. The most direct comparison is to use the ultraviolet continuum luminosities of the Ly α -selected samples and compare these with luminosity functions of the Lyman-break galaxies^{65,82} at redshifts $z \approx 3$ and 4, and at redshift $z \approx 6$ (Fig. 5; refs 15, 42). This avoids the very serious problems of estimating the dust destruction of the Ly α emission line. Here the main uncertainty in estimating a total star-formation rate is how to correct the raw counts for the fractional population with strong Ly α emission. If the estimate by Steidel *et al.*⁶⁵ based on the Lyman-break galaxies at $z \approx 3$ – 4 with strong Ly α is used to assume that the Ly α emitters represent only 20% of the star-forming galaxy population at $z \approx 5.7$, then the raw numbers for Ly α -selected galaxies (filled squares in Fig. 5) are corrected upwards to values that overlie the ultraviolet continuum luminosity functions for $z \approx 3$ and $z \approx 4$ Lyman-break galaxies. This implies that there has not been a substantial decrease in the star-formation rate out to redshift $z \approx 5.7$, in striking contrast to the direct colour break estimates. Indeed, the directly determined ultraviolet luminosity

function of the Ly α emitter sample is already comparable to the colour break determinations at $z \approx 6$ without any correction. There may be a higher fraction of Lyman α emitters at these redshifts, but it also seems likely that the current small fields used in the colour break samples are deficient in galaxies, and the rate is being underestimated when these are used.

Infrared surveys

The colour-selected and Ly α surveys can be extended to infrared wavelengths, using the same methods that have been described above. In fact, colour surveys for high-redshift galaxies applied to the GOODS fields and the HUDF already make use of infrared-continuum observations to select candidates. However, a few considerations are worth noting. For ground-based observations, the numerous strong atmospheric airglow lines at infrared wavelengths make it challenging to select dark windows for narrowband studies. Efforts⁸³ have been made to extend optical narrowband Ly α searches to the infrared, but these have not yet proved successful. A more severe problem, given the sparseness of galaxies at very high redshifts and the problems of cosmic variance, has been the limited sky area coverage for infrared detectors, typically a hundred times smaller than the areas covered in the wide-field optical surveys. This situation is changing with the availability of large-format infrared detectors and cameras.

At much higher redshifts ($z \gg 7$) the Ly α line may no longer be visible owing to the surrounding IGM, and this signature feature may not be available to us. However, it may be possible to detect the strong He II emission line at 1,640 Å associated with the first generation of metal-free Population III stars^{44,45}.

Submillimetre surveys

Submillimetre studies may be a promising route to find more-massive high-redshift galaxies. The field of imaging submillimetre galaxy studies is about 10 years old, and a class of massive, luminous dusty star-forming galaxies has been revealed^{84,85}. Such galaxies typically radiate most of their light at around 100 μm , but by reason of their enormous star-formation rate are still detected in the main submillimetre window at 850 μm , far from the peak of their emission and along the steeply falling Rayleigh–Jeans part of the black-body curve. The rapidly

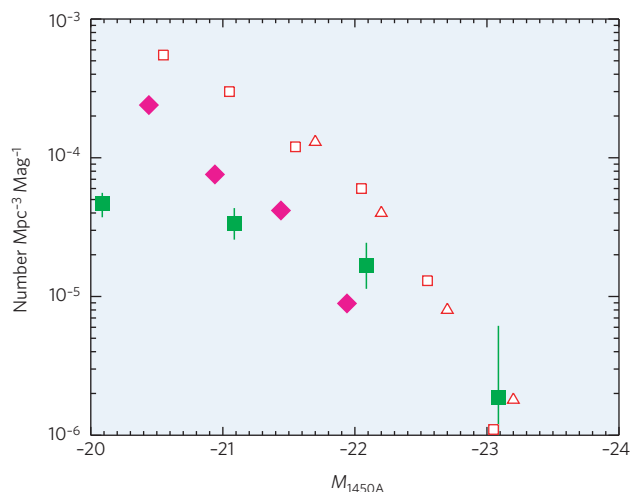


Figure 5 | Ultraviolet continuum luminosity functions for galaxies from redshifts 3 to 6. The ultraviolet continuum luminosity function of Ly α -selected $z \approx 5.7$ galaxies (green solid squares) is compared with those at $z \approx 3$ (red open squares) and $z \approx 4$ (red open triangles) reported by Steidel *et al.*⁸² and at $z \approx 6$ (solid magenta diamonds) by Bouwens *et al.*⁵². The measured points, which are taken from an area of 1.15 square degrees and are an update of Hu *et al.*⁴², are shown with 1σ Poisson uncertainties based on the number of objects in each bin; the values closely match the lower-redshift luminosity functions if emitters pick out 20% of the Lyman-break galaxy sample, as Steidel *et al.*⁶⁵ found to be the case at $z \approx 3$.

rising flux at these wavelengths counteracts the distance-dimming effects and makes these galaxies as easy to detect⁸⁶ at redshifts $z > 5$ as at $z = 1$. Most of the submillimetre light arises⁸⁷ at redshift $z \approx 1$, but a number of submillimetre sources have no obvious counterparts at any other wavelengths. It is possible that these may be at very high redshift⁸⁸. The ALMA interferometer should allow us to determine whether this is indeed the case and to measure the redshift distribution of these objects through their molecular lines.

Surveys of γ -ray bursts

The final way in which galaxies have been found at these redshifts is through the long-period γ -ray bursts (GRBs) produced in the explosions of massive stars in these galaxies. The luminosity function of the optical afterglows of γ -ray bursts extends as bright as an absolute magnitude of $M_R \approx -31.5$ mag at 1 hour after the GRB in the rest-frame, and probably even brighter at earlier epochs. Because of this extreme luminosity, they can be detected at great distances and therefore provide an exciting way to find the very high-redshift galaxies (for example see refs 9, 10) and to map the star-formation history at these extreme redshifts in a way which, if not itself unbiased, is at least largely independent of the properties of the underlying galaxies⁴⁶.

It was widely expected that the advent of the Swift satellite would produce a good rate of return⁴⁷ of very-high-redshift GRBs. Although the predictions seem, in hindsight, to have been optimistic, the Swift mission is beginning to detect GRBs at $z > 6$; GRB050904 has a spectroscopically confirmed redshift⁸⁹ of $z = 6.29$, and one further possible $z > 6$ object has recently been found. This is consistent with the star-formation histories shown in Fig. 1 or with a flat star-formation history such as might be inferred from the Lyman α emitters⁹⁰. This technique is still in its infancy but is expected to develop rapidly in the next few years.

Swift may begin to detect^{45,47} galaxies at redshifts $z > 7$ with very luminous Population III stars. This first generation of stars will have initial mass functions strongly weighted to very massive stars, and may be spectroscopically distinguished by the presence of a strong He II emission line at 1,640 Å. Such features might be confirmed with spectroscopy at infrared wavelengths, and provide evidence of a first generation of luminous stars.

Future prospects

Although there are still considerable uncertainties, we now have large samples of spectroscopically confirmed galaxies out to redshift 7 and at least a rough understanding of the star-formation history of the Universe to still higher redshifts. At least some of the objects identified at these high redshifts show strong evidence^{91–94} (dust, metals and evolved stars) that would place their first formation at still earlier times: possibly at a redshift near 20. It is clear that the frontiers of this subject will continue to move rapidly to higher redshifts in the next few years as sensitive wide-field Lyman α searches begin in the near infrared and as GRB searches begin to find yet higher redshift galaxies. We live in interesting times.

Note added in proof: The current WMAP estimate for the redshift of re-ionization is $z \approx 11$ (ref. 102).

- Gunn, J. E. & Peterson, B. A. On the density of neutral hydrogen in intergalactic space. *Astrophys. J.* **142**, 1633–1641 (1965).
- Hu, E. M. & McMahon, R. G. Detection of Lyman- α -emitting galaxies at redshift 4.55. *Nature* **382**, 231–233 (1996).
- Hu, E. M., McMahon, R. G. & Egami, E. Detection of a Ly α emission-line companion to the $z = 4.69$ QSO BR1202–0725. *Astrophys. J.* **459**, L53–L57 (1996).
- Schneider, D. P., Schmidt, M. & Gunn, J. E. PC 1247 + 3406: An optically selected quasar with a redshift of 4.897. *Astron. J.* **102**, 837–840 (1991).
- Dey, A., Spinrad, H., Stern, D., Graham, J. R. & Chaffee, F. H. A galaxy at $z = 5.34$. *Astrophys. J.* **498**, L93–L97 (1998).
- Weymann, R. J. et al. Keck spectroscopy and NICMOS photometry of a redshift $z = 5.60$ galaxy. *Astrophys. J.* **505**, L95–L98 (1998).
- Hu, E. M., McMahon, R. G. & Cowie, L. L. An extremely luminous galaxy at $z = 5.74$. *Astrophys. J.* **522**, L9–L12 (1999).
- Fan, X. et al. A survey of $z > 5.8$ quasars in the Sloan Digital Sky Survey. I. Discovery of three new quasars and the spatial density of luminous quasars at $z \sim 6$. *Astron. J.* **122**, 2833–2849 (2001).
- Hu, E. M. et al. A redshift $z = 6.56$ galaxy behind the cluster Abell 370. *Astrophys. J.* **568**, L75–L79 (2002).
- Kodaira, K. et al. The discovery of two Lyman α emitters beyond redshift 6 in the Subaru Deep Field. *Publ. Astron. Soc. Jpn* **55**, L17–L21 (2003).
- Rhoads, J. E. et al. A luminous Ly α -emitting galaxy at redshift $z = 6.535$: Discovery and spectroscopic confirmation. *Astrophys. J.* **611**, 59–67 (2004).
- Kurk, J. D. et al. A Lyman α emitter at $z = 6.5$ found with slitless spectroscopy. *Astron. Astrophys.* **422**, L13–L17 (2004).
- Taniguchi, Y. et al. The SUBARU Deep Field Project: Lyman α emitters at a redshift of 6.6. *Publ. Astron. Soc. Jpn* **57**, 165–182 (2005).
- Stern, D. et al. A galaxy at $z = 6.545$ and constraints on the epoch of reionization. *Astrophys. J.* **619**, 12–18 (2005).
- Hu, E. M., Cowie, L. L., Capak, P. & Kakazu, Y. in *IAU Colloq. 199: Probing Galaxies through Quasar Absorption Lines* (eds Williams, P. R., Shu, C. & Ménard, B.) 363–368 (Cambridge Univ. Press, Cambridge, 2005).
- Fan, X. et al. A survey of $z > 5.7$ quasars in the Sloan Digital Sky Survey. II. Discovery of three additional quasars at $z > 6$. *Astron. J.* **125**, 1649–1659 (2003).
- Kneib, J.-P., Ellis, R. S., Santos, M. R. & Richard, J. A probable $z \sim 7$ galaxy strongly lensed by the rich cluster A2218: Exploring the Dark Ages. *Astrophys. J.* **607**, 697–703 (2004).
- Bouwens, R. J. et al. Galaxies at $z \sim 7$ –8: z_{850} -dropouts in the Hubble Ultra Deep Field. *Astrophys. J.* **616**, L79–L82 (2004).
- Becker, R. H. et al. Evidence for reionization at $z \sim 6$: Detection of a Gunn–Peterson trough in a $z = 6.28$ quasar. *Astron. J.* **122**, 2850–2857 (2001).
- Fan, X. et al. Evolution of the ionizing background and the epoch of reionization from the spectra of $z \sim 6$ quasars. *Astron. J.* **123**, 1247–1257 (2002).
- Songaila, A. The evolution of the intergalactic medium transmission to redshift 6. *Astron. J.* **127**, 2598–2603 (2004).
- Fan, X. et al. Constraining the evolution of the ionizing background and the epoch of reionization with $z \sim 6$ quasars II: A sample of 19 quasars. *ArXiv Astrophysics e-prints* <astro-ph/0512082> (2006).
- Spergel, D. N. et al. First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters. *Astrophys. J. Suppl. Ser.* **148**, 175–194 (2003).
- Kogut, A. et al. First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Temperature-polarization correlation. *Astrophys. J. Suppl. Ser.* **148**, 161–173 (2003).
- Cen, R. The universe was reionized twice. *Astrophys. J.* **591**, 12–37 (2003).
- Wyithe, J. S. B. & Loeb, A. Undetected sources allow transmission of the Ly α line from galaxies prior to reionization. *Astrophys. J.* **625**, 1–5 (2005).
- Wyithe, J. S. B. & Loeb, A. Cosmic variance in the transparency of the intergalactic medium after reionization. *Astrophys. J.* (in the press).
- Haiman, Z. The detectability of high-redshift Ly α emission lines prior to the reionization of the universe. *Astrophys. J.* **576**, L1–L4 (2002).
- Furlanetto, S. R., Zaldarriaga, M. & Hernquist, L. The effects of reionization on Ly α galaxy surveys. *Mon. Not. R. Astron. Soc.* **365**, 1012–1020 (2006).
- Spinrad, H. in *Astrophysics Update* (ed. Mason, J. W.) 155–181 (Springer-Praxis, Berlin, 2004).
- Spinrad, H. *Galaxy Formation and Evolution* (Springer-Praxis, Berlin, 2005).
- Dickinson, M., Giavalisco, M. & The GOODS Team in *The Mass of Galaxies at Low and High Redshift* (eds Bender, R. & Renzini, A.) 324–331 (European Southern Observatory, 2003).
- van Breugel, W. et al. A radio galaxy at $z = 5.19$. *Astrophys. J.* **518**, L6–L64 (1999).
- Barger, A. J. et al. Very high redshift X-ray-selected active galactic nuclei in the Chandra Deep Field-North. *Astrophys. J.* **584**, L61–L64 (2003).
- Steffen, A. T. et al. An optical catalog of the Chandra Large Area Synoptic X-Ray Survey sources. *Astron. J.* **128**, 1483–1500 (2004).
- Klamer, I. J. et al. CO (1–0) and CO (5–4) observations of the most distant known radio galaxy at $z = 5.2$. *Astrophys. J.* **621**, L1–L4 (2005).
- Walter, F. et al. Molecular gas in the host galaxy of a quasar at redshift $z = 6.42$. *Nature* **424**, 406–408 (2003).
- Solomon, P. M. & Vanden Bout, P. A. Molecular gas at high redshift. *Annu. Rev. Astron. Astrophys.* **43**, 677–725 (2005).
- Miralda-Escudé, J. Reionization of the intergalactic medium and the damping wing of the Gunn–Peterson trough. *Astrophys. J.* **501**, 15–22 (1998).
- Miralda-Escudé, J. & Rees, M. J. Searching for the earliest galaxies using the Gunn–Peterson trough and the Ly α emission line. *Astrophys. J.* **497**, 21–27 (1998).
- Loeb, A. & Rybicki, G. B. Scattered Ly α radiation around sources before cosmological reionization. *Astrophys. J.* **524**, 527–535 (1999).
- Hu, E. M. et al. The luminosity function of Ly α emitters at redshift $z \sim 5.7$. *Astron. J.* **127**, 563–575 (2004).
- Hu, E. M., Cowie, L. L., Capak, P. & Kakazu, Y. Ly α emitters at redshift $z \sim 5.7$ in a wide field around the HDF-N. *Astron. J.* (submitted).
- Schaerer, D. The transition from Population III to normal galaxies: Ly α and He II emission and the ionising properties of high redshift starburst galaxies. *Astron. Astrophys.* **397**, 527–538 (2003).
- Bromm, V. & Loeb, A. High-redshift gamma-ray bursts from Population III progenitors. *Astrophys. J.* (in the press).
- Lamb, D. Q. & Reichart, D. E. Gamma-ray bursts as a probe of the very high redshift universe. *Astrophys. J.* **536**, 1–18 (2000).
- Bromm, V. & Loeb, A. The expected redshift distribution of gamma-ray bursts. *Astrophys. J.* **575**, 111–116 (2002).
- Stanway, E. R., Bunker, A. J. & McMahon, R. G. Lyman break galaxies and the star formation rate of the Universe at $z \sim 6$. *Mon. Not. R. Astron. Soc.* **342**, 439–445 (2003).
- Bunker, A. J., Stanway, E. R., Ellis, R. S. & McMahon, R. G. The star formation rate of the Universe at $z \sim 6$ from the Hubble Ultra-Deep Field. *Mon. Not. R. Astron. Soc.* **355**, 374–384 (2004).
- Bunker, A. et al. in *UC Irvine Workshop on First Light & Reionization*, *New Astron. Rev.* **50**, 94–100 (2006).
- Bouwens, R., Broadhurst, T. & Illingworth, G. Cloning dropouts: Implications for galaxy evolution at high redshift. *Astrophys. J.* **593**, 640–660 (2003).
- Bouwens, R. & Illingworth, G. Luminosity functions and star formation rates at $z \sim 6$ –10: Galaxy buildup in the reionization age. *New Astron. Rev.* **50**, 152–156 (2006).

53. Bouwens, R. J., Illingworth, G. D., Blakeslee, J. P. & Franx, M. Galaxies at $z \sim 6$: The restframe UV luminosity function and luminosity density from 506 UDF, UDF-PS, and GOODS i' -dropouts. *Astrophys. J.* (in the press).
54. Yan, H. & Windhorst, R. A. Candidates of $z \approx 5.5$ – 7 galaxies in the Hubble Space Telescope Ultra Deep Field. *Astrophys. J.* **612**, L93–L96 (2004).
55. Bunker, A. J., Stanway, E. R., Ellis, R. S., McMahon, R. G. & McCarthy, P. J. A star-forming galaxy at $z = 5.78$ in the Chandra Deep Field South. *Mon. Not. R. Astron. Soc.* **342**, L47–L51 (2003).
56. Stanway, E. R. *et al.* Hubble Space Telescope imaging and Keck spectroscopy of $z \sim 6$ i' -band dropout galaxies in the Advanced Camera for Surveys GOODS fields. *Astrophys. J.* **607**, 704–720 (2004).
57. Stanway, E. R. *et al.* Three Ly α emitters at $z \sim 6$: Early GMOS/Gemini data from the GLARE project. *Astrophys. J.* **604**, L13–L16 (2004).
58. Malhotra, S. *et al.* An overdensity of galaxies at $z = 5.9 \pm 0.2$ in the Hubble Ultra Deep Field confirmed using the ACS grism. *Astrophys. J.* **626**, 666–679 (2005).
59. Wyder, T. K. *et al.* The ultraviolet galaxy luminosity function in the local universe from GALEX data. *Astrophys. J.* **619**, L15–L18 (2005).
60. Treyer, M. A., Ellis, R. S., Milliard, B., Donas, J. & Bridges, T. J. An ultraviolet-selected galaxy redshift survey: New estimates of the local star formation rate. *Mon. Not. R. Astron. Soc.* **300**, 303–314 (1998).
61. Stiavelli, M., Fall, S. M. & Panagia, N. Observable properties of cosmological reionization sources. *Astrophys. J.* **600**, 508–519 (2004).
62. Barkana, R. & Loeb, A. Unusually large fluctuations in the statistics of galaxy formation at high redshift. *Astrophys. J.* **609**, 474–481 (2004).
63. Shimasaku, K. *et al.* Large cosmic variance in the clustering properties of Ly α emitters at $z \approx 5$. *Astrophys. J.* **605**, L93–L96 (2004).
64. Haiman, Z. & Spaans, M. Models for dusty Ly α emitters at high redshift. *Astrophys. J.* **518**, 138–144 (1999).
65. Steidel, C. C. *et al.* Ly α imaging of a proto-cluster region at $z = 3.09$. *Astrophys. J.* **532**, 170–182 (2000).
66. Rhoads, J. E. *et al.* First results from the Large-Area Lyman Alpha Survey. *Astrophys. J.* **545**, L85–L88 (2000).
67. Miyazaki, S. *et al.* Subaru prime focus camera: Suprime-Cam. *Publ. Astron. Soc. Jpn* **54**, 833–853 (2002).
68. Rhoads, J. E. *et al.* Spectroscopic confirmation of three redshift $z \approx 5.7$ Ly α emitters from the Large-Area Lyman Alpha Survey. *Astron. J.* **125**, 1006–1013 (2003).
69. Cuby, J.-G. *et al.* Discovery of a $z = 6.17$ galaxy from CFHT and VLT observations. *Astron. Astrophys.* **405**, L19–L22 (2003).
70. Dickinson, M. *et al.* Color-selected galaxies at $z \sim 6$ in the Great Observatories Origins Deep Survey. *Astrophys. J.* **600**, L99–L102 (2004).
71. Ellis, R., Santos, M. R., Kneib, J.-P. & Kuijken, K. A faint star-forming system viewed through the lensing cluster Abell 2218: First light at $z \approx 5.6$? *Astrophys. J.* **560**, L119–L122 (2001).
72. Ouchi, M. *et al.* The discovery of primeval large-scale structures with forming clusters at redshift 6. *Astrophys. J.* **620**, L1–L4 (2005).
73. Taniguchi, Y. *et al.* The discovery of a very narrow line star-forming object at a redshift of 5.66. *Astrophys. J.* **585**, L97–L100 (2003).
74. Ajiki, M. *et al.* A new high-redshift Ly α emitter: Possible superwind galaxy at $z = 5.69$. *Astrophys. J.* **576**, L25–L28 (2002).
75. Nagao, T. *et al.* A strong Ly α emitter at $z = 6.33$ in the Subaru Deep Field selected as an i' -dropout. *Astrophys. J.* **613**, L9–L12 (2004).
76. Nagao, T. *et al.* Spectroscopy of i' -dropout galaxies with an NB921-band depression in the Subaru Deep Field. *Astrophys. J.* **634**, 142–148 (2005).
77. Malhotra, S. & Rhoads, J. The volume fraction of ionized intergalactic gas at redshift $z = 6.5$. *ArXiv Astrophysics e-prints* <astro-ph/0511196> (2006).
78. Santos, M. R., Ellis, R. S., Kneib, J.-P., Richard, J. & Kuijken, K. The abundance of low-luminosity Ly α emitters at high redshift. *Astrophys. J.* **606**, 683–701 (2004).
79. Pelló, R., Schaerer, D., Richard, J., Le Borgne, J.-F. & Kneib, J.-P. ISAAC/VLT observations of a lensed galaxy at $z = 10.0$. *Astron. Astrophys.* **416**, L35–L40 (2004).
80. Bremer, M. N., Jensen, J. B., Lehnert, M. D., Förster Schreiber, N. M. & Douglas, L. Gemini H-band imaging of the field of a $z = 10$ candidate. *Astrophys. J.* **615**, L1–L4 (2004).
81. Weatherley, S. J., Warren, S. J. & Babbedge, T. S. R. Reanalysis of the spectrum of the $z = 10$ galaxy. *Astron. Astrophys.* **428**, L29–L32 (2004).
82. Steidel, C. C., Adelberger, K. L., Giavalisco, M., Dickinson, M. & Pettini, M. Lyman break galaxies at $z > 4$ and the evolution of the UV luminosity density at high redshift. *Astrophys. J.* **519**, 1–17 (1999).
83. Willis, J. P. & Courbin, F. A deep, narrow J-band search for protogalactic Ly α emission at redshifts $z \sim 9$. *Mon. Not. R. Astron. Soc.* **357**, 1348–1356 (2005).
84. Barger, A. J. *et al.* Submillimetre-wavelength detection of dusty star-forming galaxies at high redshift. *Nature* **394**, 248–251 (1998).
85. Hughes, D. H. *et al.* High-redshift star formation in the Hubble Deep Field revealed by a submillimetre-wavelength survey. *Nature* **394**, 241–247 (1998).
86. Blain, A. W. & Longair, M. S. Submillimetre cosmology. *Mon. Not. R. Astron. Soc.* **264**, 509–521 (1993).
87. Wang, W.-H., Cowie, L. L. & Barger, A. A near-infrared analysis of the submillimetre background and the cosmic star-formation history. *Astrophys. J.* (in the press).
88. Barger, A. J., Cowie, L. L. & Richards, E. A. Mapping the evolution of high-redshift dusty galaxies with submillimetre observations of a radio-selected sample. *Astron. J.* **119**, 2092–2109 (2000).
89. Kawai, N. *et al.* Afterglow spectrum of a gamma-ray burst with the highest known redshift $z = 6.295$. *Nature* (in the press).
90. Malhotra, S. & Rhoads, J. E. Luminosity functions of Ly α emitters at redshifts $z = 6.5$ and $z = 5.7$: Evidence against reionization at $z < 6.5$. *Astrophys. J.* **617**, L5–L8 (2004).
91. Egami, E. *et al.* Spitzer and Hubble Space Telescope constraints on the physical properties of the $z \sim 7$ galaxy strongly lensed by A2218. *Astrophys. J.* **618**, L5–L8 (2005).
92. Chary, R.-R., Stern, D. & Eisenhardt, P. Spitzer constraints on the $z = 6.56$ galaxy lensed by Abell 370. *Astrophys. J.* **635**, L5–L8 (2005).
93. Mobasher, B. *et al.* Evidence for a massive poststarburst galaxy at $z \sim 6.5$. *Astrophys. J.* **635**, 832–844 (2005).
94. Yan, H. *et al.* Rest-frame ultraviolet-to-optical properties of galaxies at $z \sim 6$ and $z \sim 5$ in the Hubble Ultra Deep Field: From Hubble to Spitzer. *Astrophys. J.* **634**, 109–127 (2005).
95. Schiminovich, D. *et al.* The GALEX-VVDS measurement of the evolution of the far-ultraviolet luminosity density and the cosmic star formation rate. *Astrophys. J.* **619**, L47–L50 (2005).
96. Wilson, G., Cowie, L. L., Barger, A. J. & Burke, D. J. Star formation history since $z = 1.5$ as inferred from rest-frame ultraviolet luminosity density evolution. *Astron. J.* **124**, 1258–1265 (2002).
97. Iwata, I. *et al.* Lyman break galaxies at $z \sim 5$: Luminosity function. *Publ. Astron. Soc. Jpn* **55**, 415–432 (2003).
98. Cole, S. *et al.* The 2dF galaxy redshift survey: near-infrared galaxy luminosity functions. *Mon. Not. R. Astron. Soc.* **326**, 255–273 (2001).
99. Cowie, L. L. & Hu, E. M. High- z Ly α emitters. I. A blank-field search for objects near redshift $z = 3.4$ in and around the Hubble Deep Field and the Hawaii Deep Field SSA22. *Astron. J.* **115**, 1319–1328 (1998).
100. Hu, E. M., Cowie, L. L. & McMahon, R. G. The density of Ly α emitters at very high redshift. *Astrophys. J.* **502**, L99–L103 (1998).
101. Shimasaku, K. *et al.* Subaru Deep Survey. IV. Discovery of a large-scale structure at redshift $z \approx 5$. *Astrophys. J.* **586**, L111–L114 (2003).
102. Spergel, D. N. *et al.* Wilkinson Microwave Anisotropy Probe (WMAP) three year results: implications for cosmology. *ArXiv Astrophysics e-prints* <astro-ph/0603449> (2006).

Acknowledgements This research was supported by National Science Foundation grants. We thank E. Stanway and A. Barger for their help in constructing Fig. 1.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to E.M.H. (hu@ifa.hawaii.edu).

Heavy element synthesis in the oldest stars and the early Universe

John J. Cowan¹ & Christopher Sneden²

The first stars in the Universe were probably quite different from those born today. Composed almost entirely of hydrogen and helium (plus a tiny trace of lithium), they lacked the heavier elements that determine the formation and evolution of younger stars. Although we cannot observe the very first stars — they died long ago in supernovae explosions — they created heavy elements that were incorporated into the next generation. Here we describe how observations of heavy elements in the oldest surviving stars in our Galaxy's halo help us understand the nature of the first stars — those responsible for the chemical enrichment of our Galaxy and Universe.

The Big Bang explosion at the start of the Universe was responsible for the production of the light elements hydrogen and helium, along with some fraction of lithium. All of the other elements in nature (except beryllium and boron) are synthesized inside stars, and later ejected back into the interstellar medium where they are incorporated into new generations of stars. Thus, the extinct first stars in our Galaxy and throughout the Universe were essentially composed of only hydrogen and helium — no heavy elements had yet been formed.

The chemical abundances in the oldest surviving stars, however, provide clues about the types or identities of the very first generations of stars in the Galaxy and the Universe (see recent reviews^{1–3}). Element formation is closely connected with stellar evolution. Fusion reactions (starting with hydrogen into helium) inside stars synthesize the elements up to iron. The elements heavier than iron, however, cannot be formed by fusion, and instead are synthesized as a result of slow (s-process) and rapid (r-process) neutron-capture reactions. These two nuclear processes are responsible for forming the heavier (also known as n-capture) elements such as platinum and gold. This element formation occurs during the latter stages of a star's life. Low-mass stars, similar to the Sun, live for billions of years, ending their lives as red giants and planetary nebulae. These objects are responsible for producing most of the s-process elements (such as barium), which are then ejected (at relatively low velocities) into the interstellar medium as part of the dying stars.

More massive stars evolve at a much faster rate and typically live only millions of years. During the last brief period of their lives they undergo titanic supernova explosions. It is during this explosion that the r-process elements (such as platinum and gold) are ejected into interstellar gas that will eventually form new stars. Although the general picture of element formation is understood, many questions about the nuclear physics processes and particularly the details of the supernova explosion mechanism remain to be answered^{4,5}. So the elements that are observed in the oldest stars were not synthesized internally, but instead are the result of 'seeding' from previous generations of stars. As the first generations of stars no longer exist, we suspect they must have been massive, but the details of their formation are not understood. This is particularly true because their compositions, devoid of elements except hydrogen and helium, make them different from stars like the Sun that have formed more recently. We can also tell something of the history of

star formation in our Galaxy from the iron abundance, which astronomers refer to as 'metallicity'. Most of the iron production that occurs today comes from type Ia supernovae. These result from the explosion of white dwarfs, formed from long-lived low-mass stars; thus, the stars that formed early in the history of the Galaxy and the Universe could not have had much iron. In our Galaxy these metal-poor stars are found in the (roughly spherical) halo, whereas the more metal-rich stars like the Sun reside in the flat galactic disk.

Astronomers are using the metal-poor stars in the galactic halo as laboratories for the study of n-capture element synthesis. Their chemical compositions indicate the types of synthesis processes that occurred early in the history of the Galaxy and provide clues about their progenitors, the generations of stars that lived and died before them. These abundance observations have been pushing back to lower and lower metallicities, and, presumably, earlier times. Early observational work^{6–15} focused on identifying these n-capture elements in stars with low metallicities, or $[\text{Fe}/\text{H}] < -1$. (We adopt the usual spectroscopic notations that $[A/B] \equiv \log_{10}(N_A/N_B)_* - \log_{10}(N_A/N_B)_\odot$, and that $\epsilon(A) \equiv \log_{10}(N_A/N_H) + 12.0$, for elements A and B, where N represents abundance.) More recently, very detailed abundance distributions have been obtained for the stars with metallicities $[\text{Fe}/\text{H}] \approx -3$ and large overabundances of n-capture elements, notably the stars CS22892–052 (ref. 16) and CS31082–001 (ref. 17). More such enriched stars have been identified in recent surveys^{18,19}.

The presence of these heavy elements in the observed stars demonstrates that the preceding (massive star and extinct) generation of stars in our Galaxy — perhaps the second generation of stars — was able to synthesize all of the heavy elements, before exploding as supernovae. During the past 3 years intriguing new observations^{20,21} have uncovered two stars with metallicities below $[\text{Fe}/\text{H}] = -5$. The abundance patterns in these two extremely metal-poor stars (HE0107–5240 and HE1327–2326) are rich in carbon, nitrogen and oxygen but very poor in n-capture elements, quite different from those seen in CS22892–052 and CS31082–001. These differences suggest rapid changes in the synthesis mechanisms early in the history of the Galaxy and shed light on the identity of the first stars: these were probably massive and only able to synthesize the lighter elements such as carbon, nitrogen and oxygen, not the heavy n-capture elements.

¹Homer L. Dodge Department of Physics and Astronomy, University of Oklahoma, Norman, Oklahoma 73019, USA. ²Department of Astronomy and McDonald Observatory, University of Texas, Austin, Texas 78712, USA.

The discovery of radioactive elements such as thorium and uranium in low-metallicity halo stars has led to new chronometric age determinations. Comparing the predicted and observed abundances of these long-lived radioactive elements (with precisely determined, or clock-like, half-lives) can provide a direct stellar age. This technique, originally suggested by Butcher²², has the advantage of being independent of cosmological and chemical evolution models. Thorium has now been identified in a number of stars^{13,18,23–28}. Uranium was first detected in CS31082–0129 and now is seen in several other stars^{3,18,26}. These radioactive elements have been used to date the oldest stars in our Galaxy, with typical ages of 13–14 billion years. The age estimates are in accord with determinations of globular cluster and cosmological ages, for instance from WMAP (the Wilkinson Microwave Anisotropy Probe).

In the future, planned large-scale surveys (some of which have already started) will provide new abundances and detections of very metal-poor stars, which will help in identifying the nature of the first stars and the conditions that occurred early in the history of our Galaxy. Additional high-resolution observations of spectroscopic abundances will also lead to the detections of elements never before seen in stars, as well as increasing the number of rarely seen elements such as uranium. New detections of radioactive elements such as thorium and uranium, along with physically more reliable nuclear experimental data for the heaviest, most radioactive nuclei, will also lead to more accurate dating of stellar ages. This will provide more definitive age estimates for the oldest stars, and will ultimately put strong limits on the age of our Galaxy and the Universe.

The formation of the heavy elements

Stars go through a series of fusion reactions in their interiors as they age. Through these reactions, elements from hydrogen to iron can be formed. But iron, with the largest binding energy per nucleon, cannot be fused exothermically. Additionally, as the proton number increases for heavier nuclei, the Coulomb barrier becomes too high to allow charged-particle reactions in stellar environments. So nuclei heavier than iron

are synthesized almost exclusively in neutron-capture processes (on heavy target nuclei) where there are no Coulomb barriers to overcome. These neutron-capture elements are produced either solely or by a combination of the slow and rapid processes. Our understanding of how these processes operate in stars began with pioneering work^{30,31} in 1957. In the s-process the timescale for neutron capture (τ_n) is much longer than for electron (β) decays (τ_β) from unstable nuclei. Nuclei capturing neutrons will become unstable and undergo β -decay transformations of neutrons into protons, thereby building up heavier elements. The nature of the s-process ensures that it occurs close to stability, meaning the nuclei formed are not very radioactive. The properties of this synthesis process can be determined experimentally. The situation is quite different for the r-process where $\tau_n \ll \tau_\beta$, and the nuclei produced are so unstable and short-lived that experimental information is not generally available.

Nevertheless, much can be learned about how the r-process operates. Observed elemental and isotopic abundance patterns in very old, very metal-poor stars with significant abundances of n-capture elements (see next section) agree very well with the abundance pattern for the relative or 'scaled Solar System' r-process. Happily, the one prominent success of r-process theoretical investigations has been to reproduce the Solar System abundances.

The conclusions reached for the Solar System apparently extend to all metallicities and galactic ages. One recent theoretical attempt is shown in Fig. 1 (K.-L. Kratz, B. Pfeiffer, C.S., J. W. Truran & J.J.C., manuscript in preparation), where an r-process calculation is fitted to the isotopic abundance data for the Solar System. Such calculations are intended to help us to understand the conditions, such as the temperature and neutron number densities (n_n), that are necessary for the production of these isotopes in stellar environments.

We know where s-process nucleosynthesis occurs: the sites are stars of low or intermediate mass (0.8–8 solar masses), on the asymptotic (red) giant branch, with evolutionary timescales billions of years long³². Identifying the specific astrophysical site for the r-process, however, has

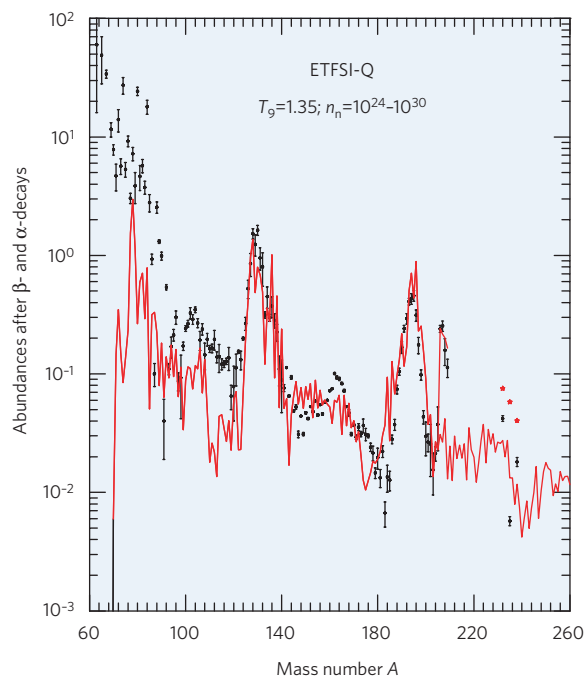


Figure 1 | Abundance comparisons between measured and predicted isotopic abundances. Dots show measured Solar System isotopic abundances (error bars indicate one standard deviation); solid line shows calculated r-process predictions. These theoretical predictions illustrate the stable abundances (after radioactive decays involving α - and β -particles), cover a range of neutron number densities, n_n , and are for a specific theoretical nuclear mass model, ETFSI-Q (K.-L. Kratz, B. Pfeiffer, C.S., J. W. Truran & J.J.C., manuscript in preparation). The temperature, T_9 , is in units of 10^9 .

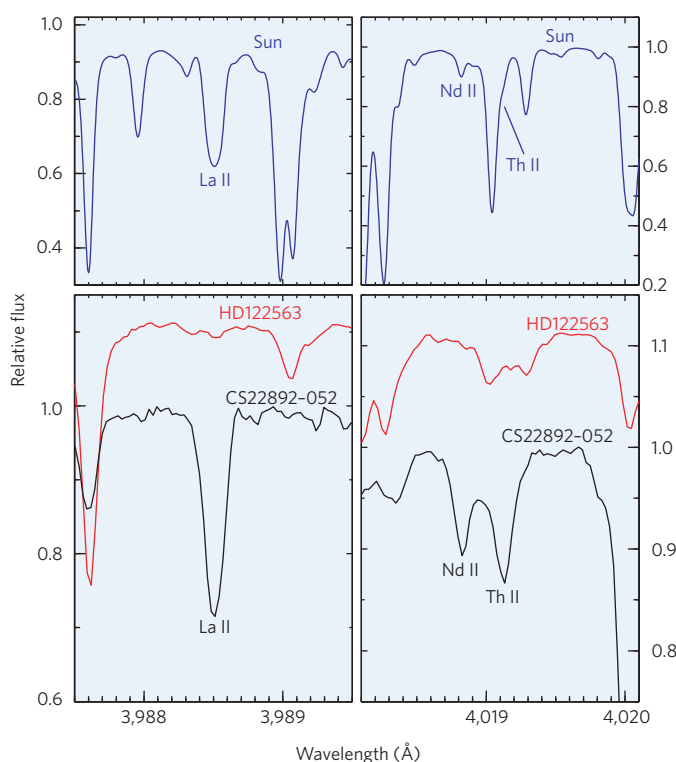


Figure 2 | Spectra featuring three n-capture lines in the Sun and two contrasting low-metallicity stars. The two low-metallicity stars are HD122563 ([Fe/H] = −2.7, [Eu/Fe] = −0.4; from ref. 24) and CS22892–052 ([Fe/H] = −3.1, [Eu/Fe] = +1.6; from ref. 16). Solar spectra are from an integrated flux atlas⁸³.

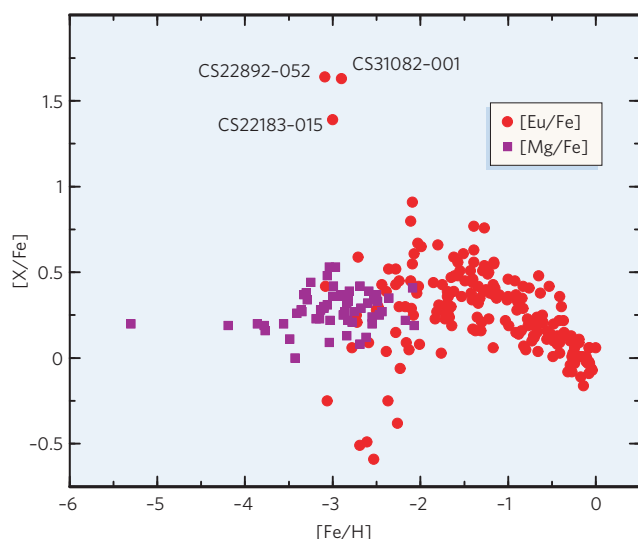


Figure 3 | Elemental ratios for galactic halo and disk stars. The plot shows abundance scatter of elemental ratios [Eu/Fe] (refs 15, 60, 63) and [Mg/Fe] (refs 64, 65) against metallicity (iron abundance) for samples of halo and disk stars in our Galaxy. The iron abundances [Fe/H] are logarithmic and with respect to solar iron. The negative numbers indicate older, metal-poor stars, whereas [Fe/H] = 0 is solar.

been much more difficult^{4,33}. The original work on this subject³⁰ suggested that the neutron-rich regions outside the nascent neutron star in a (type II) core-collapse supernova would be likely sites. More recent work has also pointed to a supernova origin, in this case the 'neutrino wind' streaming out from the central iron-core collapse of a massive star^{34,35}. Despite much work, however, the details of the supernova explosion, and the interconnected relationship with the small amount of matter associated with the r-process in this explosion, are still not well understood^{4,34,36,37}. The fact that the r-process elements are seen in some of the oldest surviving stars in the Galaxy (particularly those that are rich in heavy elements) strongly suggests that their progenitors — the very early stars that synthesized the r-process elements and ejected them explosively into the interstellar medium where they were incorporated into the halo stars — must have been rapidly evolving. Such a rapid evolution has argued against, although not necessarily ruled out, neutron-star binary mergers as an alternative (non-supernova) site for the r-process³⁸.

The r-process signatures are clearly seen in stars rich in n-capture elements in the metallicity regime [Fe/H] of approximately -2 to -3. But the two hyper-metal-poor stars discovered^{120,21} with [Fe/H] below -5 have very different abundance distributions. These stars, rich in carbon, nitrogen and oxygen with very little evidence of heavier n-capture elements, are similar in that regard to some other r-process-poor stars³⁹⁻⁴¹. These observations strongly suggest that some of the very first stars in the Galaxy, the stars that preceded the observed extremely metal-poor stars, must have only been able to synthesize elements up to iron or thereabouts. These new observations might also indicate that the stellar mass ranges of these first stars were not responsible for r-process element formation. There have been many suggestions that the earliest stellar generations in the Galaxy and the Universe were composed of very massive stars^{42,43}. For example, there have been arguments that the abundance patterns seen in HE0107-5240 and HE1327-2326 could be explained as coming from the ejecta of black-hole-forming 'faint' supernovae (of the first stars) in the mass range of 20–130 solar masses⁴⁴. The observed absence of the n-capture elements in certain of these halo stars might then be consistent with increasing evidence that r-process synthesis might be associated with lower-mass (8–10 solar mass) supernovae⁴⁵⁻⁴⁷.

Thus, the first stars might have been massive and unable to form the n-capture elements, but a somewhat later generation of early and perhaps lower-mass stars, existing before [Fe/H] = -3 to -4, could have

synthesized the total heavy-element distribution seen, for example, in CS22892-052.

Stellar abundance observations

Low-metallicity stars have widely varying relative abundance distributions. This is most apparent among the heavier n-capture elements ($Z > 55$) whose bulk contents are known to vary by two to three orders of magnitude in stars of similar metallicities. In Fig. 2 small spectral regions surrounding n-capture lines in two very metal-poor giants and the Sun are displayed. Inspection of this figure confirms the relatively simple nature of metal-poor stellar spectra, as well as the gross star-to-star variations in n-capture line strengths. Detection of thorium (strong in CS22892-052 but undetectable among the contaminating features in HD122563) is now commonplace in stars enriched in products of r-process nucleosynthesis. In Fig. 3 the star-to-star variations in n-capture elemental abundances are depicted by plotting relative abundance ratios [Eu/Fe] and [Mg/Fe] against [Fe/H] metallicity.

The abundances of the heavier n-capture elements ($Z \geq 56$) in metal-poor stars with raised abundances of r-process material closely match the Solar System r-process pattern. This is illustrated for CS22892-052 in Fig. 4. Such a clear result has only become possible recently through a combination of excellent stellar spectra and rapid gains in laboratory atomic physics. Indeed, since the publication of the CS22892-052 study, new publications^{48,49} have refined and expanded the transition data for several n-capture abundances, shrinking the uncertainties and sharpening the agreement with r-process abundances for the Solar System. This remarkable accord between r-process abundance sets created at the beginning of the Galaxy and in nucleosynthesis events leading to the formation of the Solar System strongly suggests that a unique r-process exists in nature.

Very similar abundance patterns, consistent with the solar r-abundances for the stable elements barium and above (but not necessarily including the radioactive actinides), have now been seen in essentially every other r-process-rich halo star. In Fig. 5 we show this repeated pattern from our studies and others^{17,25}. A clearly demonstrated counter-example to this statement has yet to be found. The consistency between Solar System r-process abundances and those of the r-process-rich halo stars has been extended to the isotopic level, as Sneden *et al.*⁵⁰ and Aoki *et al.*⁵¹

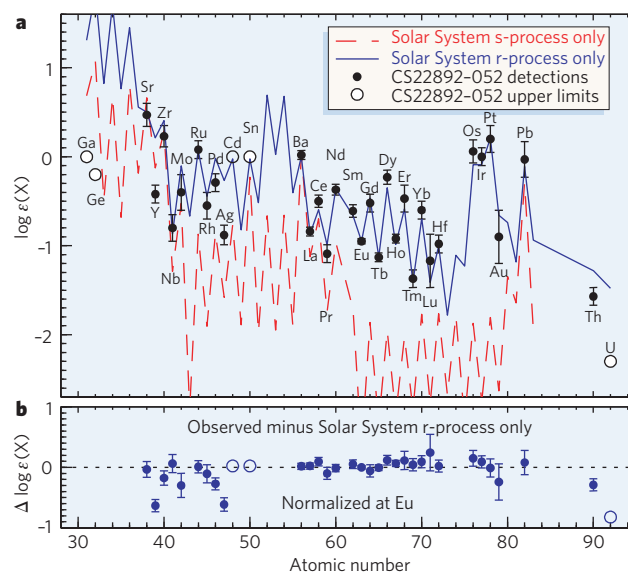


Figure 4 | Abundances in the very metal-poor halo star CS22892-052.

This figure is adapted from two figures in ref. 16. The error bars are sample deviations. **a**, Solar System r-process abundance curve⁶⁰ has been normalized to the observed Eu abundance, and the non-fitting s-process curve has been shifted so that it roughly matches the observed abundances of the Sr–Y–Zr group. **b**, Differences between the observed CS22892-052 abundances and the scaled Solar System r-process abundances are shown.

have demonstrated that $N(^{151}\text{Eu})/N(^{153}\text{Eu}) \approx 1.0 \pm 0.1$ in several halo stars; the meteoritic ratio of this r-dominant element is $0.478/0.522 = 0.92$. Additionally, Lambert and Allende Prieto⁵² have found that the ratio of odd- N isotopes to even- N ones of barium in one low-metallicity star is also consistent with a Solar System r-process-only mix.

However, inspection of Fig. 4 shows that a scaled solar r-process distribution does not seem to extend to the lighter n-capture elements. Several elements, notably silver, appear to be relatively deficient. This anomalous pattern (which also cannot be easily fitted by an abundance distribution from s-process synthesis) is found in many r-process-rich stars⁵³. Perhaps we must postulate the existence of two r-process sites^{54,55}, analogous to the two known s-processes. The formation of the lighter n-capture elements beyond the iron peak may in fact be complicated, and may require (at least in part for certain elements) some other primary synthesis process^{56,57}.

In addition to the stars within our Galaxy, there have been increasing efforts to identify abundance patterns in other galaxies. For example, metal-poor stars ($[\text{Fe}/\text{H}] < -1$) in nearby dwarf galaxies have s- and r-process abundances like those in the galactic halo⁵⁸. In addition, studies of a galaxy at high (redshift) z , with an estimated age of no more than 2.5 Gyr after the Big Bang, have detected the presence of heavy elements⁵⁹. This indicates that star formation and the related element production was rapid in galaxies that formed relatively soon after the Big Bang. This again suggests massive stars as the sites of this nucleosynthesis, during the earliest chemical history of the Universe. Even more striking in these results is that the total elemental abundance pattern seen in this galaxy, formed early in the history of the Universe — and so much older than the Sun — is consistent with a scaled Solar System abundance distribution.

Abundance trends and chemical evolution

Additional clues to the nature of the early synthesis history of the Galaxy are coming from recent studies of chemical evolution. We show in Fig. 6 the abundance trend for the element germanium as a function of metallicity, $[\text{Fe}/\text{H}]$. Germanium is normally thought of as an n-capture element with both s- and r-process contributions in Solar System material⁶⁰. Until recently data on Ge have been scarce as the dominant atomic transitions are in the ultraviolet and thus require space-based observations. The abundance behaviour shown in Fig. 6 is surprising in that the Ge abundance in these low-metallicity stars seems to track

(and is roughly proportional to) the iron abundance⁶¹. This seems to imply that Ge is formed, at least at very low metallicities and early times, in some type of charged particle reaction, or other primary process⁵⁷, that would occur in massive stars and supernovae before the s-process contribution sets in.

Other chemical evolution studies have focused on a comparison of s- and r-process elemental abundance data. These are designed to help determine the astrophysical sites for both processes and to indicate the metallicities and times for the onset of the bulk of galactic s-process nucleosynthesis. Although Ba/Eu has been used for a number of these studies, known problems in abundance determinations for barium have suggested that La/Eu might be more reliable⁶⁰. We show in Fig. 7 the variation in the ratio of La/Eu as a function of $[\text{Fe}/\text{H}]$ in a number of galactic halo and disk stars. The dashed line⁶² and the dotted line¹⁵ are predictions for the r-process-only ratio of La/Eu, and the square symbols are the values for the r-process rich stars CS22892–052, HD115444, CS31082–001 and BD+17°3248.

There are several trends apparent in this figure. First, there is a general increase in the ratio of La/Eu as the s-process contribution to lanthanum production (europium is an r-process-only element) rises with metallicity and time. This results as lower-mass stars have time to evolve, synthesize and eject s-process material that will be incorporated into new stars. Thus, the metal-rich disk stars, in general, have larger ratios than the halo stars. Second, it is seen that only the most metal-poor stars seem to have La/Eu ratios consistent with the r-process-only ratio. Although the bulk of the galactic production occurs somewhere near $[\text{Fe}/\text{H}] = -2$, Fig. 7 suggests that at least some s-process production sets in at relatively early galactic times and low metallicities^{15,60}. This clearly has implications for the mass ranges and types of stars that could have been responsible for this synthesis in the early Galaxy.

Returning to Fig. 3, we compare abundance trends as a function of metallicity for the r-process element europium and the α -element (that is, one formed by successive captures of helium nuclei, also known as α -particles) magnesium, synthesized in charged-particle reactions. What is immediately apparent is that there is a very large star-to-star scatter in the Eu/Fe abundances^{15,60,63} at low metallicities. This scatter diminishes sharply at higher metallicities. This seems to indicate that at very early times and low metallicities the Galaxy was chemically inhomogeneous in r-process elements. The same is not true for the Mg data^{64,65}, which show very little scatter at the lowest metallicities. This comparison also

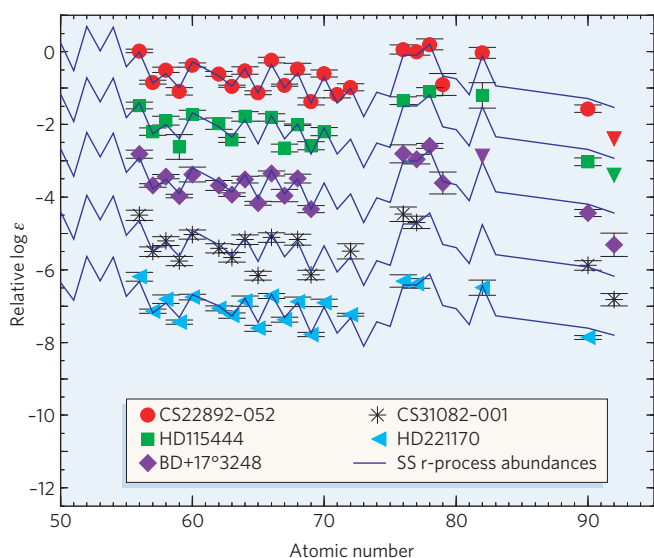


Figure 5 | Elemental abundance patterns in galactic halo stars. The plots show the neutron-capture elemental abundance pattern in the galactic halo stars CS22892–052, BD+17°3248, HD115444, CS31082–001 and HD221170 compared with the (scaled) Solar System (SS) r-process abundances (solid lines). The abundances (with sample deviation error bars) of all of the stars except CS22892–052 have been displaced downwards for display purposes.

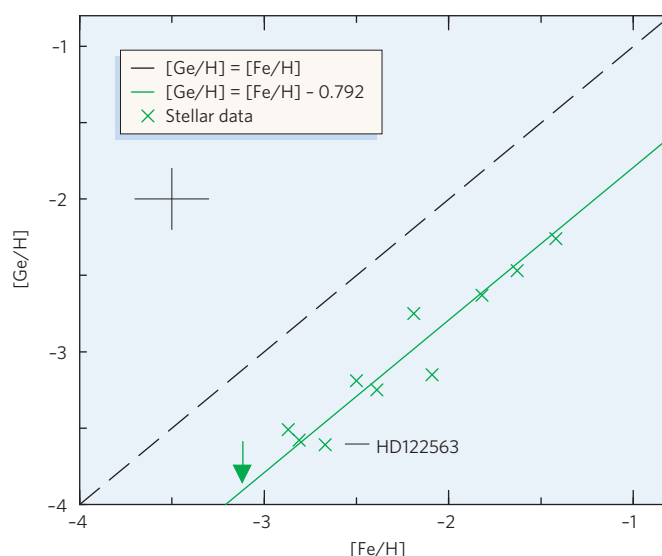


Figure 6 | Relative abundances of the elemental ratio $[\text{Ge}/\text{H}]$ as a function of metallicity (Fe). Data are for a sample of 11 galactic halo stars (after ref. 61). The arrow represents the derived upper limit for CS22892–052. The dashed line indicates the solar abundance ratio of these elements, $[\text{Ge}/\text{H}] = [\text{Fe}/\text{H}]$, and the solid green line shows the derived correlation $[\text{Ge}/\text{H}] = [\text{Fe}/\text{H}] - 0.79$. A typical error (1σ) is indicated by the cross.

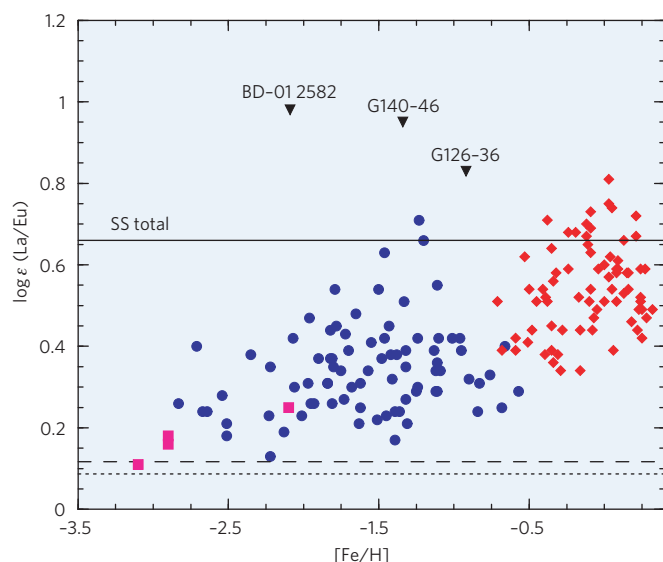


Figure 7 | Abundance trends for the ratio La/Eu in galactic stars.

Abundance trends are shown with respect to metallicity for the elemental ratio La/Eu in a large number of stars in our Galaxy (after ref. 60). The filled circles are halo stars (ref. 60), the filled diamonds are disk stars (ref. 63). The dashed (from ref. 62) and dotted (from ref. 15) lines are predictions for r-process-only ratios, and the solid line shows the total Solar System abundances. The labelled stars are all s-process rich. Pink squares represent well-studied r-process-enhanced stars.

suggests that Mg is commonly produced in most (or perhaps all) massive-star supernovae^{66,67}, whereas Eu is created more rarely in perhaps a restricted (and low-mass) supernova range.

The ages of the stars

The detection of radioactive elements such as Th, with its very long half-life of 14 Gyr (see Fig. 2), offers a promising means of determining ages. This chronometric technique depends only on determining the initial and measured abundances of these elements, usually in ratios to minimize errors. Because the low-metallicity halo stars were formed so early in the Galaxy, chemical evolutionary effects before that formation are not important. Furthermore, these stellar age estimates of the oldest stars will set a lower limit on the age of the Galaxy and hence the Universe, and these estimates will not depend on any particular cosmological model. The ideal chronometer pair is U/Th, both with long half-lives, nearby in nuclear mass and both made entirely in the r-process. The ratio U/Th has been successfully used to date CS31082-001 (12.5 ± 3 Gyr, ref. 29; 14 ± 2.4 Gyr, ref. 17; 15.5 ± 3.2 Gyr, ref. 68; 14.1 ± 2.5 Gyr, ref. 69) and BD+17°3248 (13.8 ± 4 Gyr, ref. 26). Recent estimates of the age of the Galaxy have also been made⁷⁰ by determining the production ratio of U/Th.

Uranium, however, is inherently weaker and more difficult to detect in stellar spectra than thorium. Originally, Th/Nd (ref. 22) and later Th/Eu (ref. 71) were suggested as chronometer pairs. In spite of being widely separated in mass number from Th, Eu is easily detected from ground-based observations and made almost exclusively in the r-process⁶⁰. Following the early work²³, chronometric ages based on Th/Eu have been obtained for a number of halo stars^{13,16,24-26,28,72-74}, typically with age ranges of 11–15 Gyr and with average age uncertainties of 3–4 Gyr. This approach has also yielded a similar age for one globular cluster⁷⁵. These stellar values are consistent with age determinations for the Galaxy based on main-sequence turn off ages for globular clusters^{76,77} and recent cosmological age estimates from WMAP⁷⁸ and type Ia supernovae^{79,80}.

The stellar chronometric age estimates, however, sensitively depend on the initial predicted values of Th/Eu, which, in turn, depend on the nuclear mass formulae and r-process calculations used in making those determinations (see discussion in refs 4 and 68). Typical r-process

calculations such as the one illustrated in Fig. 1 are intended to reproduce the isotopic (and elemental) abundance distribution of the Solar System. These same calculations that reproduce the stable Solar System (and stellar) elemental abundances can then be used to determine the radioactive abundances, and thus the Th/Eu ratio, at time of formation in an r-process site (a supernova). Although Th/Eu has been used in this manner for a number of stars, in at least one case (CS 31082-001) its use leads to unreasonably low age estimates. In contrast, U/Th does predict an age consistent with other ultra-metal-poor halo stars. Owing to this result and the wide separation in mass number between Th and Eu, the use of Th/Eu as a chronometer has been called into question (see discussion in ref. 28). It is not clear yet, however, whether CS31082-001 with large overabundances of Th and U is anomalous or part of a class of stars with such enhancements. We also note, however, that the reported lead abundance⁸¹, which predominantly results from α -decay of Th and U isotopes, in this star seems to be too low with respect to such high abundance values of these radioactive elements⁸². Clearly more stellar observations, particularly the detection of U, and additional theoretical and nuclear experimental studies of very neutron-rich nuclei will be required to strengthen this technique and reduce the chronometric age uncertainties.

The future

New stellar abundance observations are pushing back to lower and lower metallicities and earlier and earlier times in the history of our Galaxy. There is a concerted effort in, for example, large surveys (in progress and planned) to identify additional extremely metal-poor stars ($[\text{Fe}/\text{H}] < -5$), the generation that followed the first stars in our Galaxy. Such new observations will provide details of the synthesis histories and possible mass ranges for those first stars in our Galaxy and throughout the Universe. Additional observations will also be needed to bridge the gap between these very n-capture-poor stars at $[\text{Fe}/\text{H}] < -5$ and stars in the metallicity regime of -3 to -4 that already show complete r-process abundance signatures. Determining how the first stars exploded as supernovae will also be critical to our understanding of the nature of these early generations. The explosion mechanism (and the associated microphysics such as neutrino transport) for supernovae is still not well understood.

More work will also be required to improve our understanding of the r-process. We still, for example, do not know the astrophysical site for this process, despite decades of work. We also lack physically reliable data for the most n-rich (or radioactive) elements: perhaps new experiments such as the Radioactive Ion Accelerator, RIA, will provide this. It also remains to determine precisely the conditions that occur in the r-process, and how that affects the abundance predictions, particularly for the most radioactive nuclei. That will be especially critical in reducing the uncertainties in and allowing more precise determinations of chronometric ages. More reliable observations of the radioactive elements Th and U in the most metal-poor stars will also be needed. Together these should provide increasingly stringent constraints on estimates of the ages of our Galaxy and the Universe. ■

1. Truran, J. W., Cowan, J. J., Pilachowski, C. A. & Sneden, C. Probing the neutron-capture nucleosynthesis history of galactic matter. *Publ. Astron. Soc. Pacif.* **114**, 1293-1308 (2002).
2. Sneden, C. & Cowan, J. J. The genesis of the heaviest elements in the Milky Way galaxy. *Science* **299**, 70-75 (2003).
3. Beers, T. C. & Christlieb, N. The discovery and analysis of very metal-poor stars in the galaxy. *Annu. Rev. Astron. Astrophys.* **43**, 531-580 (2005).
4. Cowan, J. J. & Thielemann, F.-K. R-process nucleosynthesis in supernovae. *Phys. Today* **57**, 47-53 (2004).
5. Cowan, J. J. & Sneden, C. in *Carnegie Observatories Astrophysics Series, Vol. 4: Origin and Evolution of the Elements* (ed. McWilliam, A. & Rauch, M.), 27-42 (Cambridge Univ. Press, 2004).
6. Spite, M. & Spite, F. Nucleosynthesis in the Galaxy and the chemical composition of old halo stars. *Astron. Astrophys.* **67**, 23-31 (1978).
7. Sneden, C. & Parthasarathy, M. The r- and s-process nuclei in the early history of the galaxy — HD 122563. *Astrophys. J.* **267**, 757-778 (1983).
8. Sneden, C. & Pilachowski, C. A. An extremely metal-poor star with r-process overabundances. *Astrophys. J.* **288**, L55-L58 (1985).
9. Gilroy, K. K., Sneden, C., Pilachowski, C. A. & Cowan, J. J. Abundances of neutron capture elements in Population II stars. *Astrophys. J.* **327**, 298-320 (1988).

10. Gratton, R. & Sneden, C. Abundances of neutron-capture elements in metal-poor stars. *Astron. Astrophys.* **287**, 927–946 (1994).
11. McWilliam, A., Preston, G. W., Sneden, C. & Searle, L. Spectroscopic analysis of 33 of the most metal poor stars. II. *Astron. J.* **109**, 2757–2799 (1995).
12. Cowan, J. J., Burris, D. L., Sneden, C., McWilliam, A. & Preston, G. W. Evidence of heavy element nucleosynthesis early in the history of the galaxy: the ultra-metal-poor star CS22892–052. *Astrophys. J.* **439**, L51–L54 (1995).
13. Sneden, C. *et al.* The ultrametal-poor, neutron-capture-rich giant star CS22892–052. *Astrophys. J.* **467**, 819–840 (1996).
14. Ryan, S. G., Norris, J. E. & Beers, T. C. Extremely metal-poor stars. II. Elemental abundances and the early chemical enrichment of the Galaxy. *Astrophys. J.* **471**, 254–278 (1996).
15. Burris, D. L. *et al.* Neutron-capture elements in the early Galaxy: Insights from a large sample of metal-poor giants. *Astrophys. J.* **544**, 302–319 (2000).
16. Sneden, C. *et al.* The extremely metal-poor neutron capture-rich star CS22892–052: a comprehensive abundance analysis. *Astrophys. J.* **591**, 936–953 (2003).
17. Hill, V. *et al.* First stars. I. The extreme r-element rich, iron-poor halo giant CS31082–001. Implications for the r-process site(s) and radioactive cosmochronology. *Astron. Astrophys.* **387**, 560–579 (2002).
18. Christlieb, N. *et al.* The Hamburg/ESO R-process Enhanced Star survey (HERES). I. Project description, and discovery of two stars with strong enhancements of neutron-capture elements. *Astron. Astrophys.* **428**, 1027 (2004).
19. Barklem, P. S. *et al.* The Hamburg/ESO R-process enhanced star survey (HERES). II. Spectroscopic analysis of the survey sample. *Astron. Astrophys.* **439**, 129 (2005).
20. Christlieb, N. *et al.* A stellar relic from the early Milky Way. *Nature* **419**, 904–906 (2002).
21. Frebel, A. *et al.* Nucleosynthetic signatures of the first stars. *Nature* **434**, 871–873 (2005).
22. Butcher, H. R. Thorium in G-dwarf stars as a chronometer for the Galaxy. *Nature* **328**, 127–131 (1987).
23. François, P., Spite, M. & Spite, F. On the galactic age problem: determination of the Th/Eu ratio in halo stars. *Astron. Astrophys.* **274**, 821–824 (1993).
24. Westin, J., Sneden, C., Gustafsson, B. & Cowan, J. J. The r-process enriched low-metallicity giant HD 115444. *Astrophys. J.* **530**, 783–799 (2000).
25. Johnson, J. A. & Bolte, M. Th ages for metal-poor stars. *Astrophys. J.* **554**, 888–902 (2001).
26. Cowan, J. J. *et al.* The chemical composition and age of the metal-poor halo star BD+17°3248. *Astrophys. J.* **572**, 861–879 (2002).
27. Honda, S. *et al.* Spectroscopic studies of extremely metal-poor stars with the Subaru High Dispersion Spectrograph. II. The r-process elements, including thorium. *Astrophys. J.* **606**, 474–498 (2004).
28. Ivans, I. *et al.* Near-UV observations of HD 221170: new insights into the nature of r-process-rich stars. *Astrophys. J.* (in the press).
29. Cayrel, R. *et al.* Measurement of stellar age from uranium decay. *Nature* **409**, 691–692 (2001).
30. Burbidge, E. M., Burbidge, G. R., Fowler, W. A. & Hoyle, F. Synthesis of the elements in stars. *Rev. Mod. Phys.* **29**, 547–650 (1957).
31. Cameron, A. G. W. Nuclear reactions in stars and nucleosynthesis. *Publ. Astron. Soc. Pacif.* **69**, 201–222 (1957).
32. Busso, M., Gallino, R. & Wasserburg, G. J. Nucleosynthesis in asymptotic giant branch stars: relevance for galactic enrichment and Solar System formation. *Annu. Rev. Astron. Astrophys.* **37**, 239–309 (1999).
33. Cowan, J. J., Thielemann, F.-K. & Truran, J. W. The R-process and nucleochronology. *Phys. Rep.* **208**, 267–394 (1991).
34. Woosley, S. E., Wilson, J. R., Mathews, G. J., Hoffman, R. D. & Meyer, B. S. The r-process and neutrino-heated supernova ejecta. *Astrophys. J.* **433**, 229–246 (1994).
35. Takahashi, K., Witti, J. & Janka, H.-T. Nucleosynthesis in neutrino-driven winds from protoneutron stars. II. The r-process. *Astron. Astrophys.* **286**, 857–869 (1994).
36. Langanke, K. & Martínez-Pinedo, G. Nuclear weak-interaction processes in stars. *Rev. Mod. Phys.* **75**, 819–862 (2003).
37. Thompson, T. A. Magnetic proton-neutron star winds and r-process nucleosynthesis. *Astrophys. J.* **585**, L33–L36 (2003).
38. Argast, D., Samland, M., Thielemann, F.-K. & Qian, Y.-Z. Neutron star mergers versus core-collapse supernovae as dominant r-process sites in the early Galaxy. *Astron. Astrophys.* **416**, 997–1011 (2004).
39. Norris, J., Ryan, S. & Beers, T. C. Extremely metal-poor stars. VIII. High-resolution, high signal-to-noise ratio analysis of five stars with [Fe/H] < –3.5. *Astrophys. J.* **561**, 1034–1059 (2001).
40. Depagne, E. *et al.* First Stars. II. Elemental abundances in the extremely metal-poor star CS22949–037. A diagnostic of early massive supernovae. *Astron. Astrophys.* **390**, 187–198 (2002).
41. Aoki, W. *et al.* The chemical composition of the carbon-rich, extremely metal-poor star CS29498–043: a new class of extremely metal-poor stars with excesses of magnesium and silicon. *Astrophys. J.* **576**, L141–L144.
42. Qian, Y.-Z. & Wasserburg, G. Determination of nucleosynthetic yields of supernovae and very massive stars from abundances in metal-poor stars. *Astrophys. J.* **567**, 515–531 (2002).
43. Qian, Y.-Z., Sargent, W. L. W. & Wasserburg, G. The prompt inventory from very massive stars and elemental abundances in Ly α systems. *Astrophys. J.* **569**, L61–L64 (2002).
44. Iwamoto, N., Umeda, H., Tominaga, N., Nomoto, K. & Maeda, K. The first chemical enrichment in the universe and the formation of the hyper iron-poor stars. *Science* **309**, 451–453 (2005).
45. Mathews, G. J., Bazan, G. & Cowan, J. J. Evolution of heavy-element abundances as a constraint on sites for neutron-capture nucleosynthesis. *Astrophys. J.* **391**, 719–735 (1992).
46. Wheeler, C., Cowan, J. J. & Hillebrandt, W. The r-process in collapsing O/Ne/Mg cores. *Astrophys. J.* **493**, L101–L104 (1998).
47. Ishimaru, Y. & Wanajo, S. Enrichment of the r-process element europium in the galactic halo. *Astrophys. J.* **511**, L33–L36 (1999).
48. Lawler, J. E., Sneden, C. & Cowan, J. J. Improved atomic data for Ho II and new holmium abundances for the Sun and three metal-poor halo stars. *Astrophys. J.* **604**, 850–860 (2004).
49. Lawler, J. E., den Hartog, E. A., Sneden, C. & Cowan, J. J. Sm transition probabilities and abundances. *Astrophys. J.* (in the press).
50. Sneden, C. *et al.* Europium isotopic abundances in very metal-poor stars. *Astrophys. J.* **566**, L25–L28 (2002).
51. Aoki, W., Honda, S., Beers, T. C. & Sneden, C. Measurement of the europium isotope ratio for the extremely metal poor, r-process-enhanced star CS31082–001. *Astrophys. J.* **586**, 506–511 (2003).
52. Lambert, D. L. & Allende Prieto, C. The isotopic mixture of barium in the metal-poor subgiant HD140283. *Mon. Not. R. Astron. Soc.* **335**, 325–334 (2002).
53. Johnson, J. A., & Bolte, M. The r-process in the early Galaxy. *Astrophys. J.* **579**, 616–625 (2002).
54. Wasserburg, G. J., Busso, M. & Gallino, R. Abundances of actinides and short-lived nonactinides in the interstellar medium: diverse supernova sources for the r-processes. *Astrophys. J.* **466**, L109–L113 (1996).
55. Wasserburg, G. J. & Qian, Y.-Z. Prompt iron enrichment, two r-process components, and abundances in very metal-poor stars. *Astrophys. J.* **529**, L21–L24 (2000).
56. Travaglio, C. *et al.* Galactic evolution of Sr, Y and Zr: a multiplicity of nucleosynthesis processes. *Astrophys. J.* **601**, 864–884 (2004).
57. Fröhlich, C. *et al.* Neutrino-induced nucleosynthesis of A > 64 nuclei: the vp-process. Preprint astro-ph/0511376 at <http://arxiv.org/abs/> (2005).
58. Shetrone, M. *et al.* VLT/UVES abundances in four nearby dwarf spheroidal galaxies. I. Nucleosynthesis and abundance ratios. *Astron. J.* **125**, 684–706 (2003).
59. Prochaska, J. X., Howk, J. C. & Wolfe, A. M. The elemental abundance pattern in a galaxy at z = 2.626. *Nature* **423**, 57–59 (2003).
60. Simmerer, J. *et al.* The rise of the s-process in the Galaxy. *Astrophys. J.* **617**, 1091–1114 (2004).
61. Cowan, J. J. *et al.* Hubble Space Telescope observations of heavy elements in metal-poor galactic halo stars. *Astrophys. J.* **627**, 238–250 (2005).
62. O'Brien, S. *et al.* Neutron capture cross section of ¹³⁹La. *Phys. Rev. C* **68**, 035801 (2003).
63. Woolf, V. M., Tomkin, J. & Lambert, D. L. The r-process element europium in galactic disk F and G dwarf stars. *Astrophys. J.* **453**, 660–672 (1995).
64. Cayrel, R. *et al.* First stars V — abundance patterns from C to Zn and supernova yields in the early Galaxy. *Astron. Astrophys.* **416**, 1117–1138 (2004).
65. Arnone, E., Ryan, S. G., Argast, D., Norris, J. E. & Beers, T. C. Mg abundances in metal-poor halo stars as a tracer of early Galactic mixing. *Astron. Astrophys.* **430**, 507–522 (2005).
66. Woosley, S. E. & Weaver, T. A. The evolution and explosion of massive stars. II. Explosive hydrodynamics and nucleosynthesis. *Astrophys. J. Suppl.* **101**, 181–235 (1995).
67. Thielemann, F.-K. *et al.* Type II supernova nucleosynthesis and early galactic evolution. *Mem. Soc. Astron. Ital.* **71**, 483–498 (2000).
68. Schatz, H. *et al.* Thorium and uranium chronometers applied to CS31082–001. *Astrophys. J.* **579**, 626–638 (2002).
69. Wanajo, S., Itoh, N., Ishimaru, Y., Nozawa, S. & Beers, T. C. The r-process in the neutrino winds of core-collapse supernovae and U–Th cosmochronology. *Astrophys. J.* **577**, 853–865 (2002).
70. Dauphas, N. The U/Th production ratio and the age of the Milky Way from meteorites and galactic halo stars. *Nature* **435**, 1203–1205 (2005).
71. Pagel, B. E. J. in *Evolutionary Phenomena in Galaxies* (eds Beckman, J. E. & Pagel, B. E. J.) 201–223 (Cambridge Univ. Press, 1989).
72. Pfeiffer, B., Kratz, K.-L. & Thielemann, F.-K. Analysis of the solar-system r-process abundance pattern with the new EFTSI–Q mass formula. *Z. Phys. A* **357**, 235 (1997).
73. Cowan, J. J. *et al.* r-process abundances and chronometers in metal-poor stars. *Astrophys. J.* **521**, 194–205 (1999).
74. Sneden, C. *et al.* Evidence of multiple r-process sites in the early Galaxy: New observations of CS22892–052. *Astrophys. J.* **533**, L139–L142 (2000).
75. Sneden, C. *et al.* Neutron-capture element abundances in the globular cluster M15. *Astrophys. J.* **536**, L85–L88 (2000).
76. Krauss, L. M. & Chaboyer, B. Age estimates of globular clusters in the Milky Way: constraints on cosmology. *Science* **299**, 65–69 (2003).
77. Jimenez, R., Verde, L., Treu, T. & Stern, S. Constraints on the equation of state of dark energy and the Hubble constant from stellar ages and the cosmic microwave background. *Astrophys. J.* **593**, 622–629 (2003).
78. Tegmark, M. *et al.* Cosmological parameters from SDSS and WMAP. *Phys. Rev. D* **69**, 103501 (2004).
79. Riess, A. G. *et al.* Observational evidence from supernovae for an accelerating universe and a cosmological constant. **116**, 1009–1038 (1998).
80. Perlmutter, S. *et al.* Measurements of Omega and Lambda from 42 high-redshift supernovae. *Astrophys. J.* **517**, 565–586 (1999).
81. Plez, B. *et al.* Lead abundance in the uranium star CS31082–001. *Astron. Astrophys.* **428**, L9–L12 (2004).
82. Kratz, K.-L., Pfeiffer, B., Cowan, J. J. & Sneden, C. R-process chronometers. *New Astron. Rev.* **48**, 105–108 (2004).
83. Kurucz, R. L., Furenlid, I. & Brault, J. T. L. Solar flux atlas from 296 to 1300 nm. *National Solar Observatory Atlas* (National Solar Observatory, Sunspot, New Mexico, 1984).

Acknowledgements We dedicate this paper to the memory of Al Cameron, who was a giant in this field and who taught us much about nuclear astrophysics and life. We thank our colleagues for discussions and insights, and Jason Collier and Randi Worhatch for their help with the paper. We also acknowledge support from the National Science Foundation and the Space Telescope Science Institute.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to J.J.C. (cowan@nhn.ou.edu) or C.S. (chris@verdi.as.utexas.edu).

Structural basis for broad DNA-specificity in integron recombination

Douglas MacDonald¹, Gaëlle Demarre², Marie Bouvier², Didier Mazel² & Deshmukh N. Gopal¹

Lateral DNA transfer—the movement of genetic traits between bacteria—has a profound impact on genomic evolution and speciation. The efficiency with which bacteria incorporate genetic information reflects their capacity to adapt to changing environmental conditions. Integron integrases are proteins that mediate site-specific DNA recombination between a proximal primary site (*attI*) and a secondary target site (*attC*) found within mobile gene cassettes encoding resistance or virulence factors. The lack of sequence conservation among *attC* sites has led to the hypothesis that a sequence-independent structural recognition determinant must exist within *attC*. Here we report the crystal structure of an integron integrase bound to an *attC* substrate. The structure shows that DNA target site recognition and high-order synaptic assembly are not dependent on canonical DNA but on the position of two flipped-out bases that interact in *cis* and in *trans* with the integrase. These extrahelical bases, one of which is required for recombination *in vivo*, originate from folding of the bottom strand of *attC* owing to its imperfect internal dyad symmetry. The mechanism reported here supports a new paradigm for how sequence-degenerate single-stranded genetic material is recognized and exchanged between bacteria.

Nature has evolved several classes of enzymes that mediate the capture and spread of genetic information between bacteria¹. The efficiency by which these genetic traits are used is dependent on the rate of donor DNA delivery, their genomic incorporation, and the level of gene expression. Novel traits are incorporated by several methods; among these, recombination by site-specific recombinases has an important role². These enzymes require short matching DNA sequences between the donor and genomic DNA, which undoubtedly reduces the rate of DNA assimilation among highly diverse genetic populations.

Integron integrase-mediated recombination

Integron integrases (IntIs) are site-specific recombinases that form a subclass within the tyrosine recombinase family owing to the presence of a unique insertion that is required for activity³. Unlike other members of the family, such as the bacteriophage P1 Cre and yeast FLP proteins^{4,5}, IntIs can mediate the exchange of DNA between two architecturally distinct sites even though homology predicts only one DNA binding domain⁶. The mechanism by which IntIs achieve this dual site-specificity is unknown. IntIs catalyse an insertion between a primary recombination site (*attI*), located within a DNA element called an integron, and secondary target sites (*attC*), located within mobile gene cassettes⁷. These insertions are balanced by the excision of gene cassettes that occurs between two *attC* sites (Fig. 1a).

The *attI* site, like other site-specific recombinase binding sites, contains a core of short symmetrical dyad sequences at its recombination crossover point, as well as two upstream secondary sites that are potentially either regulatory or repressive towards illegitimate recombination^{8,9}. In contrast, the *attC* sites show poor sequence conservation and vary in length (57 to 141 base pairs (bp)), containing only short regions of sequence similarity at their boundaries. These conserved regions are separated by a stretch of imperfect internal dyad symmetry⁷.

It might have been expected that excision of a gene cassette would

occur through the classic model of Holliday junction (HJ) formation and resolution using two duplex *attC* sites, as observed with Cre-mediated *loxP* recombination¹⁰ (Fig. 1b). However, we have proposed an alternative pathway for IntI recombination that involves hairpin substrates formed from the bottom strand of the *attC* site, in which the HJ intermediate is resolved by as yet unknown cellular factors or perhaps DNA replication^{7,11–16} (Fig. 1c). The transition to these complex substrates requires single-stranded DNA stages that can potentially be generated through either transformation or conjugation¹⁷.

To address the structural attributes of our model, we have determined the crystal structure of IntI from *Vibrio cholerae* (VchIntIA) bound to a complex substrate (*attC*) that was derived from the bottom strand of a *V. cholerae* repeat (VCR) sequence. VCRs are the *attC* sites of superintegrations¹⁷. The structure shows that the site of recombination along the DNA backbone is determined by the position of two pre-existing extrahelical bases that not only act as molecular markers, positioning the integrase along the DNA, but also mediate the high-order assembly of the synaptic complex. The remainder of the protein–DNA interface is composed almost entirely of non-specific protein-to-DNA phosphate interactions. This structural mechanism of recognition and assembly allows a greater diversity of genetic traits to be captured and exchanged during lateral DNA transfer, thus increasing the rate of bacterial speciation.

A folded single-stranded recombination substrate

The DNA construct for co-crystallization was based on the conserved features of the predicted secondary structures of numerous *attC* bottom strands (Fig. 1d, Supplementary Fig. 1a). The resulting bulged duplex, VCR_{bs}, formed discrete complexes with VchIntIA in electrophoretic mobility shift assays (EMSAs, see below). In addition, SDS–polyacrylamide gel electrophoresis (SDS–PAGE) analysis of equilibrium mixtures of VchIntIA–VCR_{bs} complexes revealed a band with reduced mobility representing ~5% of the total protein.

¹Laboratoire de Biochimie et Biophysique des Macromolécules, Département de Biologie Structurale et Chimie, and ²Unité Postulante Plasticité du Génome Bactérien, CNRS URA 2171, Département Structure et Dynamique des Génomes, Institut Pasteur, 75724 Paris Cedex 15, France.

Mass spectrometry confirmed this slower migrating species to be covalently linked VchIntIA–VCR_{bs} molecules. This suggested the nucleophilic tyrosine (Tyr 302) within the integrase had formed the characteristic phosphoprotein intermediate.

The structure of the non-covalent VchIntIA–VCR_{bs} complex was determined to 2.8 Å using phases obtained from a single-wavelength anomalous diffraction (SAD) experiment on selenomethionine (SeMet)-labelled protein (see Methods). These phases contributed to generating a readily interpretable electron density map. Model building and refinement allowed for rapid convergence to $R_{\text{cryst}} = 0.234$ and $R_{\text{free}} = 0.262$ for 48–2.80 Å resolution data and good geometry. Data collection, phasing statistics and refinement results are summarized in Supplementary Table S1.

Architecture of the VchIntIA–VCR_{bs} excision complex

The VchIntIA–VCR_{bs} complex contains four VchIntIA molecules bound to two antiparallel VCR_{bs} duplexes (Fig. 2a). This constitutes a recombination synapse representing the step preceding first-strand cleavage in an *attC* × *attC* cassette excision reaction (Fig. 1c, step 2).

Two of the VchIntIA molecules (subunits A and C) have Tyr 302 adjacent (~3 Å) to the DNA backbone and thus attacking the scissile phosphate between nucleotides A14' and C15' on strand 1 (Supplementary Fig. 2a). The equivalent tyrosines in the B and D subunits are ~7 Å away (Supplementary Fig. 2b). The amino acids in each active site are all derived from the same subunit: that is, provided in *cis*. The extrahelical base T12'' is stabilized by *cis*-interactions within subunits B/D, and is important for DNA site recognition (see below). The extrahelical base G20'' is buried in a deep hydrophobic pocket, located in subunits A/C bound to the other VCR_{bs} duplex, thus forming a set of *trans*-interactions that hold the synaptic complex together. The protein–protein interfaces between subunits bound to the same VCR_{bs} differ from those bound in *trans* across the synapse. This arrangement yields an overall synapse that is only two-fold symmetric.

Structure of VchIntIA recombinase

VchIntIA folds into two distinct domains (Supplementary Fig. 3a). The amino-terminal domain (residues 1–85) contains four α-helices

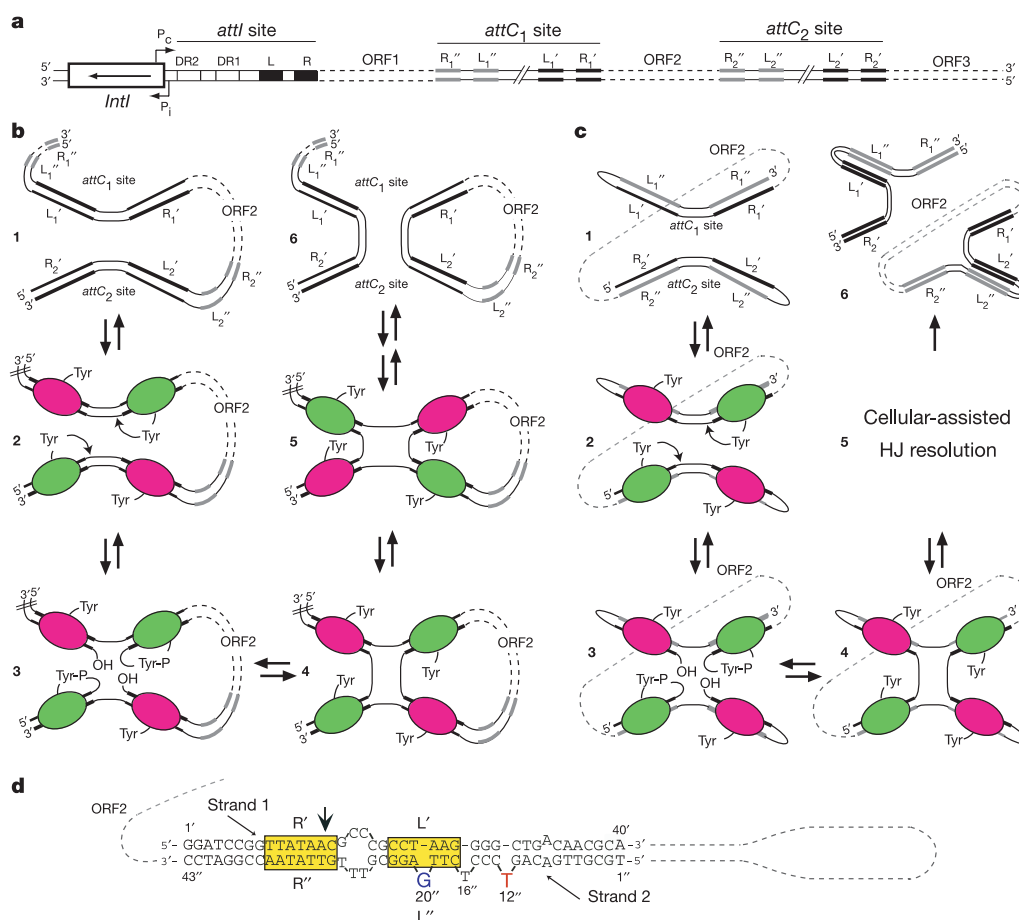


Figure 1 | Pathways of IntI-mediated cassette excision. **a**, Integrations contain a gene, *IntI*, encoding a tyrosine recombinase, and an adjacent recombination site, *attI*. Gene cassettes (open reading frames, ORFs) are flanked by secondary sites, *attC* sites. IntI recombinase *attI* and *attC* during integration and two *attC* sites during excision. *P_i* and *P_c* are promoters for *IntI* and gene cassettes, respectively; DR1 and DR2 are directly repeated accessory binding sites; L and R are binding sites within the core region of *attI*; L' and L'' are inner repeats; R' and R'' are flanking repeats. **b**, Excision by the classic tyrosine recombinase model. Each duplex *attC* site (step 1) is bound by two IntI molecules to form an antiparallel recombination synapse (step 2). Tyr 302 cleavage forms covalent 3'-phosphotyrosine intermediates (step 3). The free 5'-hydroxyl groups attack their partner substrates yielding a Holliday junction (HJ) intermediate (step 4), which isomerizes (step 5) before undergoing a second round of cleavage and strand-exchange reactions to yield the recombinant products^{5,6} (step 6). **c**, Proposed IntI

excision through a single-stranded DNA substrate pathway. The bottom strand of the integron element, produced by conjugation or transformation, folds upon itself to yield an active stem-loop substrate (step 1). Two IntI molecules bind each folded *attC* site to form an antiparallel recombination synapse (step 2). The attack and strand exchange steps proceed in a similar fashion to steps 3–4 in panel **b**; however, the HJ intermediate requires cellular components in order to be resolved¹² (steps 5–6). The reaction intermediate shown in step 2 represents the VchIntIA–VCR_{bs} structure described here. IntI molecules coloured green and magenta are potentially active or non-active for cleavage, respectively. **d**, DNA sequence of VCR_{bs} used to form VchIntIA–DNA co-crystals. Yellow boxes highlight the inner (L' and L'') and flanking (R' and R'') repeats. The nucleotides T12'' (red) and G20'' (blue) have an extrahelical geometry upon folding of *attC* bottom strands (see also Supplementary Fig. 1).

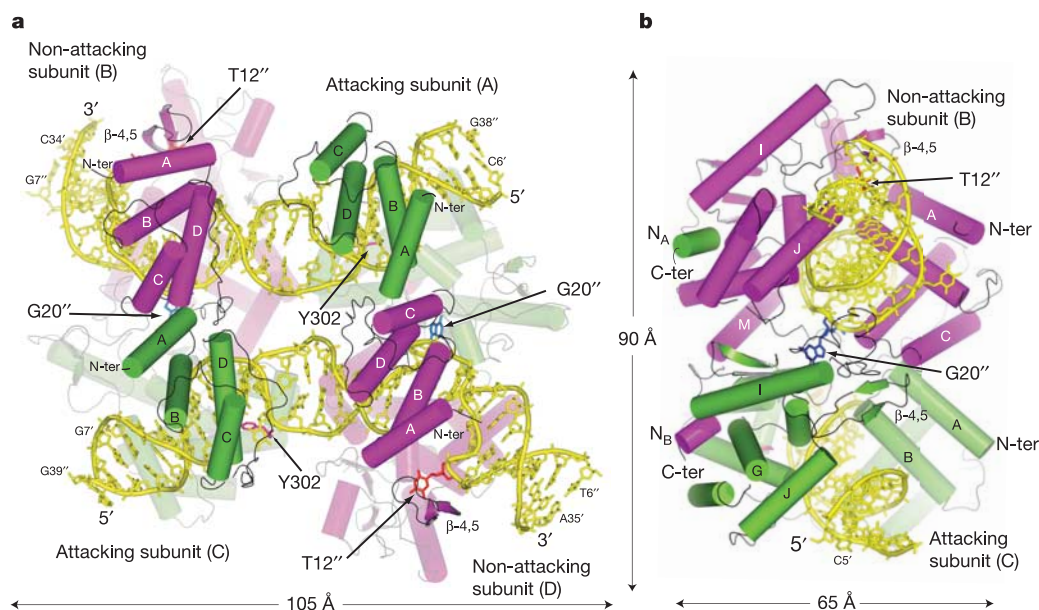


Figure 2 | Architecture of the VchIntIA-VCR_{bs} synapse. **a**, N-terminal view of the complex. Four VchIntIA molecules bind two antiparallel VCR_{bs} duplexes to form the active synapse. The extrahelical base T12'' (red) is stabilized by *cis* interactions and is involved with DNA site recognition (Fig. 4a, b). The extrahelical base G20'' (blue) is buried in subunits that are

bound to the other VCR_{bs} duplex forming a set of *trans* interactions (Fig. 4c, d). The non-symmetric interfaces between VchIntIA molecules yield a two-fold symmetric synapse. **b**, Orthogonal view with respect to **a**. The C-terminal helices (N) bury one face in a hydrophobic pocket of the adjacent subunit in a cyclic manner (N_A → B, N_B → C, and so on).

(αA–αD) organized as helix-turn-helix motifs nearly orthogonal to each other pair-wise (αA–αB, αC–αD), thus resembling the corresponding λ integrase and XerD^{18,19} folds. Helices αB and αD contact the major groove of the VCR_{bs}, whereas helix αA is involved in intersubunit contacts across the synaptic complex (Figs 2 and 3). The four separate N-terminal domains have a low root-mean-square deviation (~0.5 Å) between them.

The carboxy-terminal domain (residues 105–320) contains the characteristic insertion (residues 192–210, Supplementary Fig. 3b) found only in IntIs. The αI₂ helix within this essential region³, as discussed below, has an important role in synapse formation. The rest of the C-terminal domain is structurally similar to other tyrosine recombinase family members^{10,18,20–22}.

Substrate recognition: an adaptive molecular switch

VchIntIA binds its *attC* substrate as a dimer. However, unlike other tyrosine recombinases, it does not form symmetrical protein–DNA contacts (Supplementary Fig. 4). The attacking subunit (green, Fig. 3) positions Tyr 302 to engage the scissile phosphate, and to trigger

first-strand cleavage and transfer. The N- and C-terminal domains of this subunit wrap around the DNA, forming a clamp burying 4,000 Å² of accessible surface area. Two base contacts are made at this half-site, both through Lys 160. The positioning of this invariant catalytic residue is important, as it may have a role in the protonation of the 5'-hydroxyl-leaving group after strand cleavage²³. The remainder of this interface is characterized by protein–DNA backbone phosphate contacts.

The non-attacking integrase subunit (magenta, Fig. 3) forms two distinct protein–DNA contact points, burying 5,000 Å² of accessible surface area. One of these contacts contains the β-4,5 hairpin interacting with the flipped-out nucleotide T12'', which dictates the position of the integrase dimer along the DNA (Fig. 4a, b). This extrahelical base is inserted between two stacked histidines (His 240 and His 241; invariant among IntIs) at one end and a highly conserved proline (Pro 232) at the other end, to form a tight non-polar nucleotide–protein interface. Interruption of this interface alters the ability of VchIntIA to bind the VCR_{bs} duplex *in vitro*, as discussed below.

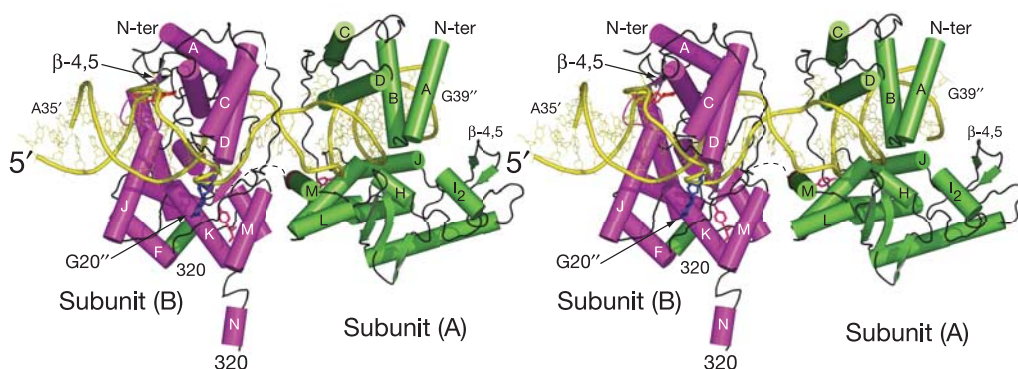


Figure 3 | Stereo-model of half the VchIntIA-VCR_{bs} synapse. The attacking VchIntIA subunit is shown in green (right) and the non-attacking VchIntIA subunit in magenta (left). Although both subunits use their N- and C-terminal domains to encompass the DNA, forming a clamp, they do

not share equivalent protein–DNA contacts (Supplementary Fig. 4). In the non-attacking subunit, the β-4,5 hairpin contacts the DNA via the extrahelical base T12''. Tyr 302 from both subunits is shown in pink.

The larger protein–DNA contact of this non-attacking subunit forms a DNA footprint that is roughly a mirror image of the attacking subunit interface, covering half a helical turn of DNA and sharing many of the equivalent non-specific protein–DNA phosphate contacts (Supplementary Fig. 4). Notably, this subunit—although centred on G20''—does not make direct contacts with this extrahelical base. Instead, this flipped-out base binds within a deep hydrophobic pocket located in the attacking subunit (C) across the synaptic interface (Fig. 2b).

IntIs must recognize two architecturally distinct sites during cassette integration ($attI \times attC$). To achieve this dual site-specificity, we propose that IntIs have developed a molecular switch that allows sequence-degenerate binding ($attC$), as seen in the VchIntIA–VCR_{bs} complex reported here, as well as sequence-dependent binding during $attI$ site recognition. The molecular switch that adapts to these two different modes of binding is the β -4,5 hairpin. During VCR_{bs} ($attC$) binding, the β -4,5 hairpin is found in two environments corresponding to each unique half-site along the DNA. In the non-attacking subunit, the β -4,5 hairpin recognizes the molecular marker T12'' (Fig. 4a, b), whereas in the attacking one it does not contact the DNA (Fig. 4c). We suggest that in the attacking interface, the position of the β -4,5 hairpin may have rotated away from the minor groove relative to its position when IntIs are in complex with an $attI$ site, possibly owing to the *trans*-interaction mediated by G20'' (Supplementary Fig. 5). During $attI$ binding, the β -4,5 hairpin would make contacts with the minor groove analogous to the equivalent disposition of this element in the Cre–*lox* system¹⁰. The increased DNA footprint for $attI$ site binding (Supplementary Fig. 4) relative to the $attC$ interface reported here supports such a hypothesis⁸.

Synaptic assembly: a multiple structure problem

Assembly of a tyrosine recombinase synapse is mediated by a highly specific set of intersubunit protein–protein interactions, induced by DNA target capture. This stepwise assembly ensures that strand exchange takes place only between appropriate DNA substrates. How can IntIs, which bind a vast array of $attC$ sequences, guarantee competent assembly of their DNA excision synapse? The strategy they have adopted is to use an invariant flipped-out DNA base, acting as a linchpin, to position the two dimers in the synapse (Fig. 4c, d). This flipped-out base (G20'') guarantees proper geometrical assembly by inserting itself within the hydrophobic pocket, created by Trp 157 and Trp 219, across the synaptic interface. This interaction allows helix αI_2 from the attacking subunit(s) to form several important DNA contacts in *trans*, holding the synapse together.

These *trans* interactions mediated by G20'' result in only a two-fold symmetric synapse, which may inhibit HJ isomerization as discussed below. Subunits bound to the same VCR_{bs} bury only 1,600 Å² of solvent-accessible protein surface, by docking the C-terminal helix (αN) of the attacking subunit in the neighbouring non-attacking subunit^{10,24} (Fig. 2b, Fig. 3). The spacing and geometry of the bases in the central region of the DNA duplex preclude intersubunit interaction between the N termini on the same VCR_{bs} (Fig. 2a, Fig. 3). The more extensive intersubunit interface across the synapse buries 4,000 Å² of accessible surface area. These interfacial contacts are mostly due to C-terminal helix (αN) exchange, helix αA in the attacking subunits contacting the helix-turn-helix region of the αC – αD helices within the non-attacking subunits across the synapse, and the unusual *trans* interactions mediated by the extrahelical base G20''. A Supplementary Movie showing the final model of the VchIntIA–VCR_{bs} synaptic complex is available in Supplementary Information.

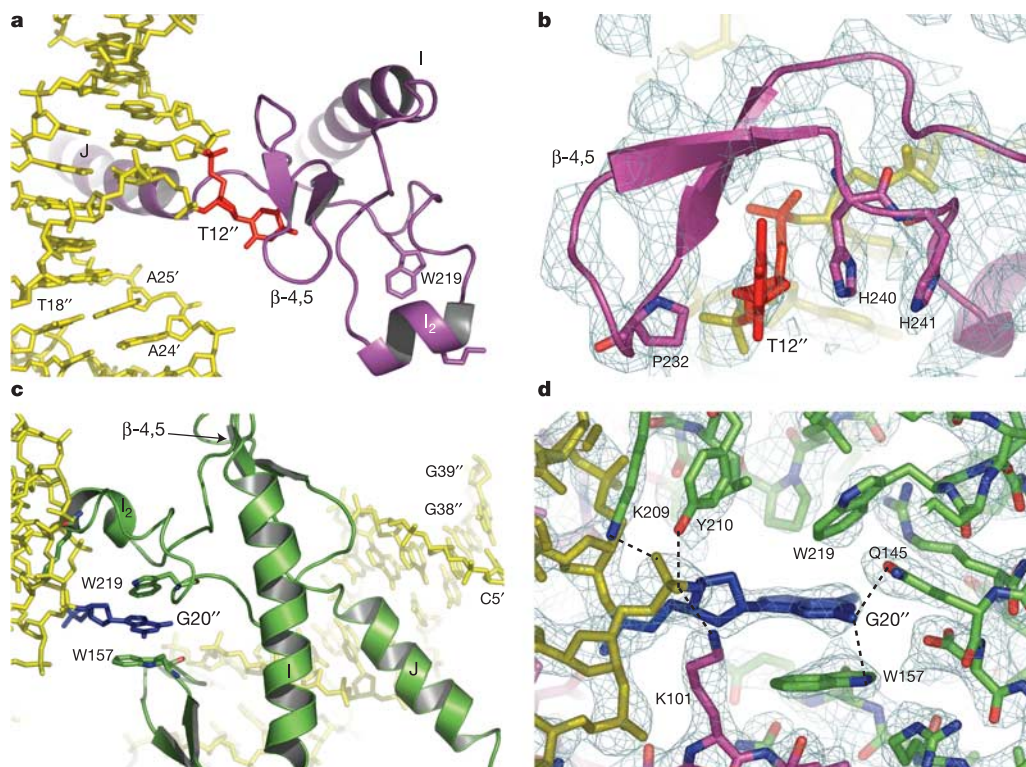


Figure 4 | *Cis* and *trans* extrahelical interactions. **a**, Interaction with β -4,5 from a non-attacking subunit and the extrahelical base T12''. **b**, Detailed view of the *cis* interaction depicted in **a**. Base T12'' (red) forms a protein–nucleic acid interface by stacking between His 240 and Pro 232. Experimental electron densities after phase modification are shown (1.5 σ contour level). **c**, Ribbon diagram showing G20'' (blue) binding in *trans* between Trp 157 and Trp 219. The β -4,5 hairpin from the attacking subunit

does not contact the DNA. **d**, Detailed view of the *trans* interactions shown in **c**. The NH₂ group of G20'' forms hydrogen bonds with the O γ of Glu 145 and N1 of Trp 157 within the attacking subunits. Non-specific contacts are made with the adjoining non-attacking subunit (magenta). A 2.8-Å 2F_o – F_c simulated annealed omit electron density map (1.6 σ contour level) is shown. G20'' and all protein residues labelled in **d** were omitted during map calculation.

Integron site-specific recombination

Recently, exploiting conjugation as a medium to exclusively deliver single-stranded *attC* substrates, we demonstrated that the bottom strand of *attC* recombined with a resident *attI* site at a rate that was 1,000-fold higher than the comparable top strand of *attC*¹². Disruption of the postulated^{13–16} secondary structure of *attC* affected recombination. On the basis of these results we proposed a recombination model (Fig. 1c). The structure of the VchIntIA–VCR_{bs} complex presented here provides a structural basis for IntI-mediated site-specific recombination using the bottom strand of *attC* as a substrate.

The biological relevance of the observed structural aspects of our model—that is, the *cis* and *trans* interactions observed in the VchIntIA–VCR_{bs} complex—were investigated using a combination of EMSAs and *in vivo* excision assays¹³. Four discrete bands are observed with wild-type VchIntIA–VCR_{bs} (Supplementary Fig. 6a), potentially corresponding to the occupancy of the four available half-sites. Deletion of T12'' from VCR_{bs} markedly reduced the overall amount of complex formation as judged by EMSA (Supplementary Fig. 6a), with, however, only a fivefold drop in excision frequency (Supplementary Fig. 6c). Mutagenesis of His 240 (H240V), which interacts with T12'' in the crystal, did not allow us to further investigate this interaction as the mutant protein formed aggregates with VCR_{bs} (Supplementary Fig. 6b, lane 5). Notably, deletion of G20'' from VCR_{bs} affected both the band intensity and the migration pattern, with a quantitative increase of the fastest migrating band (Supplementary Fig. 6a). This suggests a higher occupancy of only one half-site and a change in the effective radius/molecular weight or nature of higher-order complexes. This same mutated sequence tested *in vivo* resulted in a 10,000-fold drop in excision activity (Supplementary Fig. 6c). Preliminary *in vivo* mutagenesis experiments of key residues (W157I and W219I) also produced a large reduction (1,000-fold) in excision frequency. These results suggest that G20'' has a central role in maintaining an active VchIntIA–VCR_{bs} synapse, but that other factors—namely C-terminal helix exchange, and B/C and A/D interfaces—also contribute to the assembly of the tetrameric synapse.

A folded single-strand *attC* substrate during integron recombination necessitates that second-strand cleavage and transfer is down-regulated relative to most members of the tyrosine recombinase family. A second round of cleavage and transfer does not lead to the excision/insertion of the gene cassettes but only to rearrangements within the *attC* sites (Supplementary Fig. 7). This regulation probably occurs at the HJ isomerization step, and hence perturbs the next stage of the reaction in IntI recombination²⁵. Studies based on λ , XerC/D, Flp and Cre systems support models of HJ isomerization that involve only limited branch migration resulting in subtle movements of the quaternary structure^{10,20,26–28}. The two isomeric HJ intermediates have structures that are similar except for an exchange in the roles of their DNA strands²⁹. In this model, the protein–protein interface between subunits is maintained. Small changes between subunits, which result in the loss of interfacial binding energy, are regained by reciprocal changes in symmetrically related interfaces. These HJ intermediates thus have similar free energies, which explain the lack of bias observed in their resolution in the Cre and Flp systems³⁰. These reciprocal intersubunit changes require a pseudo four-fold symmetric synapse. We suggest that as the VchIntIA–VCR_{bs} synapse is only two-fold symmetric—and thus contains non-equivalent subunit interfaces (Fig. 2a)—the reciprocal changes needed to maintain the iso-energetic intermediates are lost. This may lead to an increased energy barrier for isomerization and thus produce a larger population of HJ intermediates that have not isomerized—that is, with a bias to revert to the original substrates. However, this group of stalled HJ intermediates then needs to be resolved via other cellular processes (Fig. 1c).

In addition, the rotation ($\sim 15^\circ$) of the C-terminal domains within

the non-attacking subunits could also reduce the rate of second-strand cleavage (Supplementary Fig. 8). This movement is a consequence of binding the extrahelical base T12''. It results in the translation of helix α M, and repositions nucleophilic Tyr 302 away from the DNA backbone to a distance of ~ 7 Å in the non-attacking subunits.

Increased genomic diversity through broad DNA-specificity

Several mechanisms mediate genetic exchange between prokaryotes, but the introduction of foreign DNA does not guarantee its assimilation. Processes such as DNA recombination are needed if persistence as an episome is not established. These recombination systems, however, place restrictions on the type of DNA that can be transferred. Homologous recombination primarily rearranges sequences among closely related taxa and is unlikely to allow for the introduction of novel traits³¹. Site-specific recombination, by contrast, can mediate the exchange of more diverse traits, but normally these systems are still constrained by the need for the necessary core sites for recognition and catalysis². However, IntIs have evolved an ingenious way of allowing highly diverse genetic traits, containing little sequence homology at their DNA target sites, to be recognized and assimilated into the integron element. Although the positioning of DNA bases in an extrahelical geometry for genomic repair is well documented³², the use of extrahelical bases for DNA recognition of single-stranded DNA, as well as assembly of higher-order protein–DNA complexes, is gathering more support^{12,33,34}. We suspect that as most of the DNA exchanged through lateral transfer is in the single-stranded form, be it through natural transformation or conjugation³⁵, this may have contributed to the development of the mechanism reported here.

METHODS

Structure determination and refinement. The preparation of VchIntIA and the DNA used to form co-crystals is described in Supplementary Information. X-ray diffraction data to 2.8 Å was collected at beamline ID14-4 of the European Synchrotron Radiation Facility (ESRF). A wavelength of 0.9793 Å, corresponding to the SeMet peak, was used for data collection at 100 K on a single co-crystal. X-ray data were processed with XDS³⁶ and CCP4³⁷ program suites. Phases for the VchIntIA–VCR_{bs} complex were determined by using the single-wavelength anomalous dispersion (SAD) method with SeMet-labelled protein. The position of substructure atoms, initial phases and solvent flattening were determined using BnP³⁸. The atomic model was fitted into electron density using the program O³⁹. The asymmetric unit contained four VchIntIA molecules and two antiparallel VCR_{bs} duplexes. Residues 1 and 305–310 in subunit A and 305–307 in subunit C were not fit into the density. For the DNA helices, nucleotides 1–4 and 36–40 of chain E, nucleotides 1–5 and 40–43 of chain F, nucleotides 1–5 and 35–40 of chain G, and nucleotides 1–6 and 39–43 of chain H were not fit owing to poor electron density. DNA chains E and G correspond to strand 1, and DNA chains F and H correspond to strand 2 in Fig. 1d. The four protein subunits and DNA duplexes were all independently modelled and refined with CNS⁴⁰ using maximum-likelihood target using amplitudes and phase probability distribution with alternate cycles of manual density fitting. The structural similarity of subunits A/C and B/D warranted the use of non-crystallographic symmetry restraints in the final rounds of refinement. This resulted in an $R_{\text{cryst}} = 0.234$ and $R_{\text{free}} = 0.262$ at 2.8 Å, with good geometry for bond lengths and angles. All amino acids have (ϕ , ψ) backbone torsion angles in allowed regions of Ramachandran space. Statistics for the final model are given in Supplementary Table S1. The figures were produced with PYMOL (<http://www.pymol.org>).

VCR_{bs} and VchIntIA mutagenic analysis. Details of EMSAs and preparation of mutants is described in Supplementary Information. The *in vivo* excision assay has been previously described¹³. Briefly, the assay measures the frequency at which a synthetic integron gene cassette—the reporter gene (*lacI*^q)—is excised from its two flanking recombination sites, *attC*_{aadA7} and the VCR_{2/1}. In all experiments, symmetrical mutations were introduced to both the VCR_{2/1} and *attC*_{aadA7} sites when possible. The equivalent base to T16'' is not present within *attC*_{aadA7} (see Supplementary Fig. 1a). This protocol was followed as previous experiments showed mutations introduced in either site independently led to a slight decrease in deletion frequency, whereas mutations introduced at identical positions in tandem resulted in considerably larger reductions in recombination

frequency. This suggests that the presence of a wild-type *attC* site assists in integrase complex formation between a mutated site, probably in a cooperative fashion. This observation may be similar to the formation of the *attC* × *attI* synapse, where the extrahelical base G20^o within the *attC* site drives formation of the active synapse (D. Mazel, unpublished observations). Mutations in the *attC*_{aadA7} and VCR_{2/1} sites were introduced as described previously¹³.

Received 29 October 2005; accepted 10 February 2006.

- Kazazian, H. H. Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
- Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. (eds) *Mobile DNA II* (ASM Press, Washington DC, 2002).
- Messier, N. & Roy, P. H. Integron integrases possess a unique additional domain necessary for activity. *J. Bacteriol.* **183**, 6699–6706 (2001).
- Chen, Y. & Rice, P. A. New insight into site-specific recombination from Flp recombinase–DNA structures. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 135–159 (2003).
- Van Duyne, G. A structural view of tyrosine recombinase site-specific recombination. In *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) (ASM Press, Washington DC, 2002).
- Azaro, M. A. & Landy, A. Lambda integrase and the lambda Int family. In *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) (ASM Press, Washington DC, 2002).
- Stokes, H. W., O’Gorman, D. B., Recchia, G. D., Parsekhian, M. & Hall, R. M. Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol. Microbiol.* **26**, 731–745 (1997).
- Gravel, A., Fournier, B. & Roy, P. H. DNA complexes obtained with the integrase IntI1 at the *attI1* site. *Nucleic Acids Res.* **26**, 4347–4355 (1998).
- Stark, W. M., Boocock, M. R. & Sherratt, D. J. Site-specific recombination by Tn3 resolvase. *Trends Genet.* **5**, 304–309 (1989).
- Guo, F., Gopaul, D. N. & van Duyne, G. D. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* **389**, 40–46 (1997).
- Hansson, K., Skold, O. & Sundstrom, L. Non-palindromic *attI* sites of integrons are capable of site-specific recombination with one another and with secondary targets. *Mol. Microbiol.* **26**, 441–453 (1997).
- Bouvier, M., Demarre, G. & Mazel, D. Integron cassette insertion: a recombination process involving a folded single strand substrate. *EMBO J.* **24**, 4356–4367 (2005).
- Biskri, L., Bouvier, M., Guerout, A. M., Boissard, S. & Mazel, D. Comparative study of class 1 integron and *Vibrio cholerae* superintegron integrase activities. *J. Bacteriol.* **187**, 1740–1750 (2005).
- Rowe-Magnus, D. A., Guerout, A. M., Biskri, L., Bouige, P. & Mazel, D. Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. *Genome Res.* **13**, 428–442 (2003).
- Johansson, C., Kamali-Moghaddam, M. & Sundstrom, L. Integron integrase binds to bulged hairpin DNA. *Nucleic Acids Res.* **32**, 4033–4043 (2004).
- Francia, M. V., Zabala, J. C., de la Cruz, F. & Garcia Lobo, J. M. The IntI1 integron integrase preferentially binds single-stranded DNA of the *attC* site. *J. Bacteriol.* **181**, 6844–6849 (1999).
- Mazel, D., Dychinco, B., Webb, V. A. & Davies, J. A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280**, 605–608 (1998).
- Subramanya, H. S. et al. Crystal structure of the site-specific recombinase, XerD. *EMBO J.* **16**, 5178–5187 (1997).
- Aihara, H., Kwon, H. J., Nunes-Duby, S. E., Landy, A. & Ellenberger, T. A conformational switch controls the DNA cleavage activity of λ integrase. *Mol. Cell* **12**, 187–198 (2003).
- Chen, Y., Narendra, U., Iype, L. E., Cox, M. M. & Rice, P. A. Crystal structure of a Flp recombinase–Holliday junction complex: assembly of an active oligomer by helix swapping. *Mol. Cell* **6**, 885–897 (2000).
- Kwon, H. J., Tirumalai, R., Landy, A. & Ellenberger, T. Flexibility in DNA recombination: structure of the lambda integrase catalytic core. *Science* **276**, 126–131 (1997).
- Hickman, A. B., Waninger, S., Scocca, J. J. & Dyda, F. Molecular organization in site-specific recombination: the catalytic domain of bacteriophage HP1 integrase at 2.7 Å resolution. *Cell* **89**, 227–237 (1997).
- Krogh, B. O. & Shuman, S. Catalytic mechanism of DNA topoisomerase IB. *Mol. Cell* **5**, 1035–1041 (2000).
- Rafferty, J. B. et al. Crystal structure of DNA recombination protein RuvA and a model for its binding to the Holliday junction. *Science* **274**, 415–421 (1996).
- Gopaul, D. N. & Van Duyne, G. D. Structure and mechanism in site-specific recombination. *Curr. Opin. Struct. Biol.* **9**, 14–20 (1999).
- Arciszewska, L. K., Grainge, I. & Sherratt, D. J. Action of site-specific recombinases XerC and XerD on tethered Holliday junctions. *EMBO J.* **16**, 3731–3743 (1997).
- Biswas, T. et al. A structural basis for allosteric control of DNA recombination by λ integrase. *Nature* **435**, 1059–1066 (2005).
- Nunes-Duby, S. E., Azaro, M. A. & Landy, A. Swapping DNA strands and sensing homology without branch migration in λ site-specific recombination. *Curr. Biol.* **5**, 139–148 (1995).
- Gopaul, D. N., Guo, F. & Van Duyne, G. D. Structure of the Holliday junction intermediate in Cre–loxP site-specific recombination. *EMBO J.* **17**, 4175–4187 (1998).
- Meyer-Leon, L., Inman, R. B. & Cox, M. M. Characterization of Holliday structures in FLP protein-promoted site-specific recombination. *Mol. Cell. Biol.* **10**, 235–242 (1990).
- Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
- Klimasauskas, S., Kumar, S., Roberts, R. J. & Cheng, X. HhaI methyltransferase flips its target base out of the DNA helix. *Cell* **76**, 357–369 (1994).
- Val, M. E. et al. The single-stranded genome of phage CTX is the form used for integration into the genome of *Vibrio cholerae*. *Mol. Cell* **19**, 559–566 (2005).
- Ronning, D. R. et al. Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Mol. Cell* **20**, 143–154 (2005).
- Llosa, M., Gomis-Ruth, F. X., Coll, M. & de la Cruz, F. Bacterial conjugation: a two-step mechanism for DNA transport. *Mol. Microbiol.* **45**, 1–8 (2002).
- Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26**, 795–800 (1993).
- Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Weeks, C. M. et al. Towards automated protein structure determination: BnP, the SnB-PHASES interface. *Z. Kristallogr.* **217**, 686–693 (2002).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brunger, A. T. et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by laboratory start-up funds from the Pasteur Institute to D.N.G., research grants from the CNRS to D.N.G., research grants from the Institut Pasteur and from the CNRS on transposable elements to D. Mazel, and a French Ministry of Research (MENESR) grant to D. Mazel and D.N.G. D. MacDonald is a Florence Gould Scholar funded by the Pasteur Foundation of New York. G.D. and M.B. are doctoral fellows from the Fondation pour la Recherche Médicale and MENESR, respectively. We also thank T. Ellenberger and G. Liou for critical review of this manuscript.

Author Contributions D. MacDonald carried out the *in vitro* assays, structural determination and wrote the paper with editing from D.N.G. D. Mazel provided the conceptual framework for the substrate and invaluable background. G.D. and M.B. performed the *in vivo* excision frequency assays. All authors discussed the results and commented on the manuscript.

Author Information Coordinates and structure factors are deposited in the Protein Data Bank under accession code 2A3V. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.N.G. (gopauld@pasteur.fr).

Origin of the obliquities of the giant planets in mutual interactions in the early Solar System

Adrián Brunini^{1,2}

The origin of the spin-axis orientations (obliquities) of the giant planets is a fundamental issue because if the obliquities resulted from tangential collisions with primordial Earth-sized protoplanets, then they are related to the masses of the largest planetesimals out of which the planets form^{1–7}. A problem with this mechanism, however, is that the orbital planes of regular satellites would probably be uncorrelated with the obliquities, contrary to observations^{6,7}. Alternatively, they could have come from an external twist that affected the orientation of the Solar System plane²; but in this model, the outer planets must have formed too rapidly, before the event that produced the twist⁸. Moreover, the model cannot be quantitatively tested. Here I show that the present obliquities of the giant planets were probably achieved when Jupiter and Saturn crossed the 1:2 orbital resonance⁹ during a specific migration process: different migration scenarios cannot account for the large observed obliquities. The existence of the regular satellites of the giant planets does not represent a problem in this model because, although they formed soon after the planetary formation, they can follow the slow evolution of the equatorial plane it produces.

During the process of accumulation, planets acquire rotational angular momentum from the relative motions at collision of the material from which they accrete^{3–7}. Owing to the symmetry of the problem about the plane of the planet's orbit, a spin angular momentum either in the same direction as the orbital angular momentum (prograde rotation) or in the opposite direction (retrograde rotation) is possible.

The systematic component of planetary rotation does not account for the spin obliquity (the angle between the planet equator and the orbital plane) of the giant planets⁵. Nevertheless, the stochastic nature of planetary accretion from planetesimals allows for a random component to the vector spin in any direction⁶. Because planets might accumulate a significant fraction of their mass and angular momentum from only a few large impacts, stochastic effects may be important in determining spin obliquities¹. Thus the large spin obliquity of the giant planets, particularly that of Uranus (98°), was attributed to a tangential collision with an Earth-sized protoplanet^{7,10,11}.

Because of their equatorial bulge, the planets are subject to torques from the gravitational forces of their satellites, the other planets and the Sun. This causes a spin-axis precession. Moreover, the obliquity of each planet suffers perturbations due to the secular motion of the planet's orbit. The inner planets could have started with nearly zero obliquity, and the chaotic behaviour of their obliquities could have driven them to their current values¹², but the obliquities of the giant planets are essentially stable¹².

Nevertheless, the orbital architecture of the Solar System in the past could have been very different from today's. Planetary migration is a natural process during planet formation¹³. After a short phase in which the giant planets acquired most of their mass from the

surrounding gas, they became immersed in a residual disk of small planetesimals. The exchange of angular momentum with these planetesimals caused the outer planets to migrate far from their birthplaces. Recent numerical simulations⁹ assuming an originally compact outer planetary system were able, for the first time, to reproduce the main orbital features of the outer Solar System, provided Jupiter and Saturn crossed the 1:2 orbital resonance during their migration.

The resonance crossing excites the orbital eccentricities of the outer planets and, if the system is initially rather compact, a short phase of planetary encounters starts, drastically affecting the further orbital evolution. As the obliquities of the giant planets may have been strongly influenced by the mutual interactions during this phase, I have designed a series of numerical simulations of the evolution of the obliquities of the giant planets during the migratory stage, assuming a compact initial structure of the outer Solar System⁹. I have incorporated the radial migration by means of a tangential force, as in ref. 14, with its magnitude tuned to represent the radial migration induced by a planetesimal disk extended from just outside the outer planet's initial orbit, and up to 30 AU (where 1 AU is the Sun–Earth distance) as in ref. 9. I have also modelled the effect of the disk of planetesimals on the orbital eccentricities and inclinations of the planets, with an eccentricity damping timescale consistent with the theory of dynamical friction^{15,16}.

As the best match with the present orbital architecture of the outer Solar System was obtained in ref. 9 when the initial orbits of the four giant planets allow Saturn to experience encounters with one of the ice giants, I restrict my runs to this class of initial configurations (labelled 'class B' runs in ref. 9).

One of the main quantities determining the evolution of the obliquity of an oblate planet is its precession constant, which depends on a number of physical and orbital parameters⁸. I have adopted the present values for all of them. In addition, I considered that a large fraction of the external torques is exerted on the satellite systems of the planet, rather than directly on the planet itself⁸. My model includes the torque exerted by the Sun, and the mutual torques from the four giant planets. The rotational equations were simultaneously integrated with the orbital evolution. In all the runs, the initial obliquity of each planet was set arbitrarily small.

In total I have performed 30 runs, differing only in the timescale of migration (1–10 million years) and in the initial semimajor axes of the planets (chosen at random, but following class B runs as in ref. 9; see Supplementary Information). In 19 of the runs I have obtained successful outer Solar System analogues. The orbital evolution of one representative run is shown in Fig. 1. For the remaining 11 runs, in five runs one of the ice giants was quickly ejected from the Solar System, and in six runs, the final orbital configuration was inconsistent with the present Solar System, and were not included in the statistical analysis.

¹Facultad de Ciencias Astronómicas y Geofísicas, Universidad Nacional de La Plata, Paseo del Bosque s/n (1900) La Plata, Argentina. ²Instituto Astrofísico de La Plata, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Paseo del Bosque s/n (1900) La Plata, Argentina.

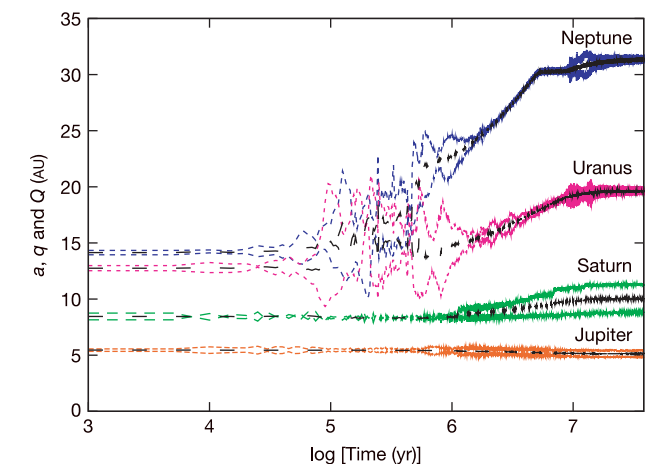


Figure 1 | The orbital evolution of the giant planets in one of the simulations that match the orbital configuration of the outer Solar System. Three curves are plotted for each planet: the semimajor axis (*a*), and the minimum (*q*) and maximum (*Q*) heliocentric distances. The separation between the curves is representative of the orbital eccentricity. In the first stages, Saturn and Jupiter cross the 1:2 orbital resonance, producing strong changes in the orbits of the ice giants, which approach one another and eventually Saturn. The orbits were circularized as though dynamical friction by the planetesimal disk were operating while the planets were immersed in the planetesimal disk (the disk is assumed to start just beyond the initial orbit of the outermost planet). The integration of the orbital and rotational equations was performed with a high-precision Bullirsch–Stoer routine²³.

I found that the obliquity evolution is strongly influenced by the exchange of angular momentum during planetary encounters. In particular, Uranus and Saturn acquire large obliquities during the first phase of planetary migration, during which they usually approach one another. (In the six runs discarded, Uranus also acquires a large obliquity.) Neptune’s obliquity increases by the action of the mutual Uranus–Neptune perturbations. I observed that the largest obliquities are produced when Neptune starts its evolution inside the orbit of Uranus (~50% of the runs). Jupiter did not experience large obliquity variations. Nevertheless it acquires a non-negligible obliquity. The bulk statistics of the results are summarized in Table 1.

I made one additional run to extract from the model the external torques due to the mutual perturbations (leaving only the solar torque) and the obliquity evolution was rather regular. The same occurred in ~60 additional runs carried out before the publication of the compact model⁹, with an initial system much less compact¹³.

The gaseous nebula, whose presence could delay the resonance crossing event¹⁷, has little influence on the obliquities¹⁸. Assuming that such a delay occurred, my initial conditions should be considered the dynamical state of the giant planets immediately after the dissipation of the nebula.

One of the main problems with the big impact hypothesis is that the regular satellites of the giant planets orbit in the equatorial plane of the planets, in the same direction as they spin. There are strong arguments suggesting that these satellites formed near the equatorial plane very soon after, or simultaneously during, the planetary

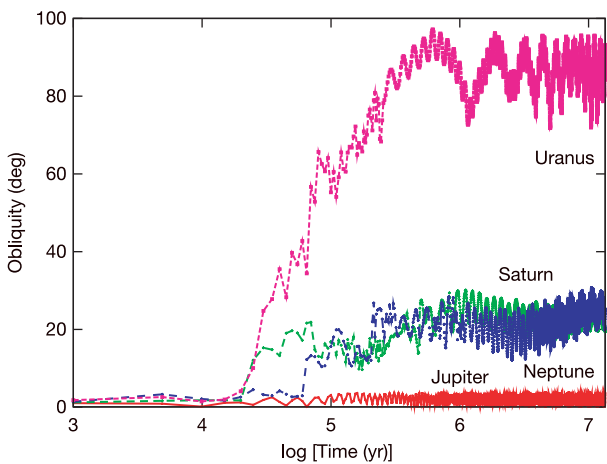


Figure 2 | The evolution of the spin obliquity of the giant planets. Each planet is represented by the same colour as in Fig. 1. Originally, all the obliquities are small (~1°) and then growth mainly occurs during a short phase after the passage of Jupiter and Saturn through the mutual 2:1 orbital resonance. In this period, planetary encounters excite the obliquities up to nearly the observed values. The most affected planet is, on average, Uranus, that, in contrast to Neptune, also suffers encounters with Saturn. This happens only in models of an originally compact outer planetary system. Less compact initial configurations do not bring originally small obliquities to such high values. After the initial period, once the planets are far from one another, the further obliquity evolution is regular. Note that large libration amplitudes are naturally acquired in this model.

formation¹⁹. However, it is not possible that they existed before a big collision that could tilt the planetary spin axis, because there is no efficient mechanism known to be able to move the satellite orbital plane back to the planetary equator.

Because the planets are oblate, the nodes of the satellite orbits precess around the equatorial plane of the planets. If this rate of precession is faster than the rate of change of the obliquity, as is the case for the present planetary precession rate, it is possible for the satellites to follow the changes in the spin-axis orientation, always remaining in the equatorial plane²⁰. The longest precession period of the nodes of the main regular satellites of Uranus (the most critical case) is ~260 years. It is much shorter than the timescale of the typical evolution of the obliquity (see Fig. 2). Nevertheless, I also integrated the orbits of Uranus’ five regular satellites according to the oblateness perturbations, and changed Uranus’ obliquity according to the outputs of five of my ‘successful’ runs. In all cases, the satellites remained near the plane of the planetary equator, following the changes in obliquity. I conclude that the presence of the natural satellites does not represent a problem in this scenario.

Recently⁸, a resonant interaction between the precession period of Saturn’s spin axis and the precession period of Neptune’s orbit was claimed to be responsible for tilting Saturn’s spin axis. This explanation, however, does not seem to extend to the ice giants. My model has the advantage of explaining the origin of the obliquities of the four giant planets through an unified mechanism, and, in contrast to the ‘twist’ theory², it was possible to present a quantitative test.

Table 1 | Bulk statistics of the simulations’ results

Planet	Semimajor axis, <i>a</i> (AU)	Eccentricity, <i>e</i>	Inclination, <i>i</i> (deg)	Obliquity, <i>ε</i> (deg)
Jupiter	5.2 ± 0.02 [5.2]	0.03 ± 0.006 [0.05]	1.5 ± 0.2 [1.3]	2.9 ± 0.3 [3.1]
Saturn	9.7 ± 0.1 [9.5]	0.08 ± 0.01 [0.05]	2.2 ± 0.2 [2.5]	22.4 ± 8.3 [25.6]
Uranus	21.4 ± 0.5 [19.3]	0.01 ± 0.005 [0.04]	1.6 ± 0.7 [0.8]	87.2 ± 6.6 [97.8]
Neptune	30.4 ± 0.1 [30.3]	0.007 ± 0.002 [0.008]	2.0 ± 0.8 [1.8]	29.6 ± 4.3 [28.3]

Average values of the orbital elements and obliquities at the end of the 19 runs that best match the present orbital configuration of the outer Solar System. They are average values taken over the last five million years of each simulation. The actual values are shown in square brackets and the errors refer to standard deviations.

My model generates the giant planet obliquities from mutual interactions, and so it predicts that the total rotational angular momentum vector of the planets should remain close to the orbital angular momentum vector. I note that in the true Solar System, the total rotational angular momentum vector is shifted with respect to the orbital angular momentum by less than 5° . My results add support to a primitive compact outer Solar System, in which Jupiter and Saturn crossed the 1:2 orbital resonance⁹ during a migration process, and Uranus and Neptune formed closer to the gaseous giants.

The existence of a large population of primordial Earth-sized planetoids in the Uranus–Neptune region has been invoked to explain a number of dynamical features observed in the populations of small bodies in the Solar System²¹, but they are difficult to reconcile with modern results of planetary accretion²². Nothing much larger than Pluto could form in a standard model of protoplanetary solar nebula. Such small planetoids do not suffice to tilt the spin axes of the giant planets to their present values nor to explain the orbital excitation in the asteroid belt and in the Kuiper belt. My results suggest that other alternatives to explain these observed facts should be found.

Received 7 October 2005; accepted 2 January 2006.

1. Safronov, V. S. *Evolution of the Protoplanetary Cloud and Formation of The Earth and the Planets* Ch 10, 113–135 (NASA, 1969).
2. Tremaine, S. On the origin of the obliquities of the outer planets. *Icarus* **89**, 85–92 (1991).
3. Dones, L. & Tremaine, S. On the origin of planetary spins. *Icarus* **103**, 67–92 (1993).
4. Lissauer, J. J., Berman, A. F., Greenzweig, Y. & Kary, D. M. Accretion of mass and spin angular momentum by a planet on an eccentric orbit. *Icarus* **127**, 65–92 (1997).
5. Lissauer, J. J. & Kary, D. M. The origin of the systematic component of planetary rotation. *Icarus* **94**, 126–159 (1991).
6. Lissauer, J. J. & Safronov, V. S. The random component of planetary rotation. *Icarus* **93**, 288–297 (1991).
7. Brunini, A. A possible constraint to Uranus' great collision. *Planet. Space Sci.* **43**, 1019–1021 (1995).
8. Ward, W. & Hamilton, D. P. Tilting Saturn I. Analytic model. *Astron. J.* **128**, 2501–2509 (2004).
9. Tsiganis, K., Gomes, R., Morbidelli, A. & Levison, H. F. Origin of the orbital architecture of the giant planets of the Solar System. *Nature* **435**, 459–461 (2005).
10. Korycansky, D. G., Bodenheimer, P., Cassen, P. & Pollak, J. B. One-dimensional calculations of a large impact on Uranus. *Icarus* **84**, 528–541 (1990).
11. Parisi, M. G. & Brunini, A. Constraints to Uranus' great collision. II. *Planet. Space Sci.* **45**, 181–187 (1997).
12. Laskar, J. & Robutel, P. The chaotic obliquity of the planets. *Nature* **361**, 608–612 (1993).
13. Fernández, J. A. & Ip, W.-H. Some dynamical aspects of the accretion of Uranus and Neptune—The exchange of orbital angular momentum with planetesimals. *Icarus* **58**, 109–120 (1984).
14. Malhotra, R. The origin of Pluto's peculiar orbit. *Nature* **365**, 819–821 (1993).
15. Ward, W. R. On disk–planet interactions and orbital eccentricities. *Icarus* **73**, 330–348 (1988).
16. Artymowicz, P. Disk–satellite interaction via density waves and the eccentricity evolution of bodies embedded in disks. *Astrophys. J.* **419**, 166–180 (1993).
17. Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature* **435**, 466–469 (2005).
18. Miki, S. The gaseous flow around a protoplanet in the primitive solar nebula. *Prog. Theor. Phys.* **67**, 1053–1067 (1982).
19. Colombo, G. & Franklin, F. A. On the formation of the outer satellite groups of Jupiter. *Icarus* **15**, 186–189 (1971).
20. Goldreich, P. Inclination of satellite orbits about an oblate precessing planet. *Astron. J.* **70**, 5–9 (1965).
21. Petit, J.-M., Morbidelli, A. & Valsecchi, G. B. Large scattered planetesimals and the excitation of the small body belts. *Icarus* **141**, 367–387 (1999).
22. Thommes, E. W., Duncan, M. J. & Levison, H. F. Oligarchic growth of giant planets. *Icarus* **161**, 431–455 (2003).
23. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. *Numerical Recipes: The Art of Scientific Computing* Ch 16, 718–726 (Cambridge Univ. Press, New York, 1989).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the support of IALP and ANPCyT.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to the author (abrunini@fcaglp.unlp.edu.ar).

LETTERS

Wave and defect dynamics in nonlinear photonic quasicrystals

Barak Freedman¹, Guy Bartal¹, Mordechai Segev¹, Ron Lifshitz², Demetrios N. Christodoulides³
& Jason W. Fleischer^{4,5}

Quasicrystals are unique structures with long-range order but no periodicity. Their properties have intrigued scientists ever since their discovery¹ and initial theoretical analysis^{2,3}. The lack of periodicity excludes the possibility of describing quasicrystal structures with well-established analytical tools, including common notions like Brillouin zones and Bloch's theorem. New and unique features such as fractal-like band structures^{4–7} and 'phason' degrees of freedom⁸ are introduced. In general, it is very difficult to directly observe the evolution of electronic waves in solid-state atomic quasicrystals, or the dynamics of the structure itself. Here we use optical induction^{9–11} to create two-dimensional photonic quasicrystals, whose macroscopic nature allows us to explore wave transport phenomena. We demonstrate that light launched at different quasicrystal sites travels through the lattice in a way equivalent to quantum tunnelling of electrons in a quasiperiodic potential. At high intensity, lattice solitons are formed. Finally, we directly observe dislocation dynamics when crystal sites are allowed to interact with each other. Our experimental results apply not only to photonics, but also to other quasiperiodic systems such as matter waves in quasiperiodic traps¹², generic pattern-forming systems as in parametrically excited surface waves¹³, liquid quasicrystals¹⁴, and the more familiar atomic quasicrystals.

We produce two-dimensional photonic quasicrystals by using the optical induction technique^{9,10} that has recently become a useful tool in studying solitons in two-dimensional nonlinear optical lattices¹¹. A similar technique was used in the past to trap cold atoms in quasiperiodic potentials¹². Other theoretical¹⁵ and experimental techniques, based on direct fabrication at the level of individual structural elements, have been used for creating photonic quasicrystals for the general study of their optical properties¹⁶, nonlinear frequency conversion applications^{17,18}, or the generation of complete photonic bandgap materials¹⁹. The recent discovery of self-assembled dendrimer liquid quasicrystals¹⁴ may lead to another interesting method, but optical induction is currently the only approach that allows for full structural control and real-time manipulation of the generated photonic quasicrystals.

The optical induction technique relies on the interference of several monochromatic light beams, whereby the resulting intensity pattern is translated into a (periodic or quasiperiodic) change in the refractive index of a photosensitive nonlinear material, in our case photorefractive SBN:75 ($\text{Sr}_{0.75}\text{Ba}_{0.25}\text{Nb}_2\text{O}_6$, see refs 9, 11). For photonic quasiperiodic structures, we overlap five coherent monochromatic laser beams with wave vectors $\mathbf{k}_1, \dots, \mathbf{k}_5$ separated by angles of $2\pi/5$. The experimental image of the field-intensity pattern is shown in Fig. 1a. It consists of a d.c. $\mathbf{k} = 0$ component and exactly 20 harmonic components with wave vectors of the form $\mathbf{k}_i - \mathbf{k}_j$

($i, j = 1 \dots 5$), as indicated by the appearance of 20 Bragg peaks in the calculated diffraction image (Fig. 1c). We emphasize that the induced refractive index change in our material is of a saturable nature. We intentionally work in the saturated regime, so the induced photonic quasicrystal contains additional higher-harmonic Bragg peaks, as clearly visible in the experimental diffraction diagram (Fig. 1b). Evidently, the refractive index structure impressed into the volume of the nonlinear material is a fully-fledged photonic quasicrystal. From the relative phases of these Fourier components, we infer that the induced quasicrystal has decagonal (ten-fold, as opposed to only pentagonal) symmetry. Furthermore, one can tile the induced decagonal pattern using rhombic Penrose tiles, producing a

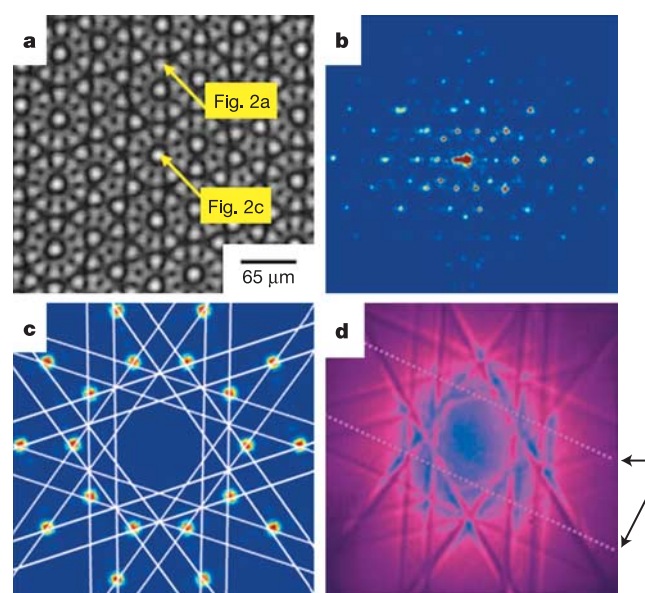


Figure 1 | Photonic quasicrystal and its extended Brillouin zone map.

a, Experimental image of the decagonal field-intensity pattern generated by five writing beams with wave vectors $\mathbf{k}_1, \dots, \mathbf{k}_5$ separated by angles of $2\pi/5$. (Yellow arrows mark the probe beam excitation sites for Fig. 2.) **b**, Experimental diffraction pattern showing additional higher-harmonic Bragg peaks, confirming the creation of a fully-fledged photonic quasicrystal. **c**, **d**, Theory (**c**) and experiment (**d**) showing the effective Brillouin zones of the optically induced decagonal quasicrystal. Panel **c** shows the 20 Bragg peaks (after subtraction of the $\mathbf{k} = 0$ component) corresponding to all wave vectors of the form $\mathbf{k}_i - \mathbf{k}_j$ ($i, j = 1 \dots 5$) that make the real space intensity patterns. Arrows at right of **d** indicate lines appearing 'weak' in the experiment. Bright colours indicate higher intensity.

¹Physics Department and Solid State Institute, Technion-Israel Institute of Technology, Haifa 32000, Israel. ²School of Physics & Astronomy, Tel Aviv University, Tel Aviv 69978, Israel. ³College of Optics and Photonics, University of Central Florida, Florida 32816-2700, USA. ⁴Electrical Engineering Department, ⁵Princeton Institute for the Science and Technology of Materials, Princeton University, New Jersey 08544, USA.

decagonal tiling that contains ten-fold star configurations, sometimes referred to as the ‘anti-Penrose’ tiling.

There are two fundamentally different types of lattices that can be created using the optical induction method: fixed and deformable. In the first case, the induced array is essentially rigid, while in the second case the individual sites can move and interact with each other. Technically, these two cases differ in the way the writing beams are polarized, but the end result is far more important. The photo-refractive crystal has the property that arbitrarily polarized light induces changes in its refractive index according to the field intensity. On the other hand, only light that is polarized in a certain direction (extraordinary polarization in our SBN:75 crystal) experiences a spatially-varying index of refraction. Light that is polarized in the orthogonal direction (ordinary polarization in SBN:75) propagates through the medium as if its refractive index were homogeneous. The first optical induction scheme involves ordinarily polarized writing beams that induce the quasicrystal but do not interact with each other, and an additional extraordinarily polarized probe beam^{9–11}. The quasicrystal produced in this way is ideally suited to studying the linear and nonlinear tunnelling transport of a wavepacket in a fixed quasiperiodic lattice (Fig. 2), in a way similar to previous studies in square and hexagonal lattices¹¹. The second scheme of optical induction uses extraordinarily polarized writing beams, which are affected by the induced changes in the refractive index and therefore

interact with each other as the crystal is induced, giving rise to complex nonlinear dynamics²⁰. That is, the sites interact with one another through the optical nonlinearity, in a way similar to the interactions between an array of solitons in homogeneous media. This kind of induced crystal can be used for the study of dynamical processes such as defect healing, dislocation motion, and phason strain relaxation (Figs 3 and 4). (An explanation of the term ‘phason’ appears below.)

We start our experiments by mapping the extended effective Brillouin zones (Jones zones) of the induced photonic structure (Fig. 1c, d). This is done by employing the recently developed Brillouin zone spectroscopy method²¹, which relies on visualizing the power spectra of the optical Bloch modes²² propagating through the photonic crystal. Similar results, showing the effective Brillouin zones of an icosahedral photonic quasicrystal, yet via very many point-by-point measurements, have been reported recently¹⁶. The excellent agreement between the experimental momentum-space picture (Fig. 1d) and Bragg diffraction peaks (Fig. 1b), and the expected effective Brillouin zones predicted by the positions of the Bragg peaks (Fig. 1c), proves not only that what we have induced is a photonic quasicrystal, but also that this quasicrystal is capable of carrying optical Bloch modes, extended along its axis of ten-fold symmetry.

The standard benchmark experiment to study wave dynamics in lattices is ‘discrete diffraction’, which is the optical equivalent of quantum tunnelling in a periodic potential^{23,24}. A focused probe beam is launched into a single site and experiences lateral transport within the lattice as it propagates along its axis, resulting in a diffraction pattern characteristic of the local structure that is probed^{11,25}. In a simple periodic crystal, with a single ‘atom’ per unit cell, each site looks like any other and discrete diffraction looks the same irrespective of the initial site of excitation. In quasicrystals, however, there are always a number of different local environments, so that transport behaviour is expected to vary significantly from site to site. We observe this behaviour in the decagonal photonic quasicrystal numerically (Fig. 2a, c) and experimentally (Fig. 2b, d), by comparing the discrete diffraction from two different sites (marked in Fig. 1a). The two points chosen differ from one another by their local environments, one having a higher local symmetry than the other, as clearly observed in the numerical and experimental linear propagation (discrete diffraction) images.

Unlike this linear phenomenon of discrete diffraction, in many physical systems (such as photonic quasicrystals and matter waves in a quasiperiodic potential) the underlying quasiperiodic structure behaves nonlinearly to waves of large-enough amplitude, and can support lattice solitons (discrete solitons)^{9–11,23–25}. We therefore conducted a series of experiments (Fig. 2e, f), and observed nonlinear localization of the discrete diffraction pattern. In this set of experiments, we used a positive bias field, hence the nonlinearity is of the self-focusing type¹¹. At sufficiently high intensity of the probe beam (while keeping the photonic quasicrystal unchanged), we obtained the first observation of lattice solitons in a quasiperiodic waveguide structure (Fig. 2f). To verify the robustness of the lattice soliton, we conducted several more nonlinear localization experiments, by changing the intensity of the probe beam while keeping all other parameters fixed; that is, by varying the intensity ratio of the soliton/lattice from 1:3 to 1:4. Throughout this range, the output intensity distribution of the soliton remains unchanged (Fig. 2f). We have also tested the robustness of the soliton to the lattice depth, by varying the bias field. The soliton is robust to such variations within $\sim 5\%$. Altogether, soliton formation and propagation is a robust phenomenon in the quasicrystal, with an experimentally large basin of attraction.

We now move on to study the dynamics of the photonic quasicrystal, using interacting extraordinarily polarized beams, so it is essential to maintain a stable ground-state structure of the interacting sites of the crystal. Such behaviour, of a stable interacting square lattice of the self-focusing type (where all the lattice sites are the same), was recently

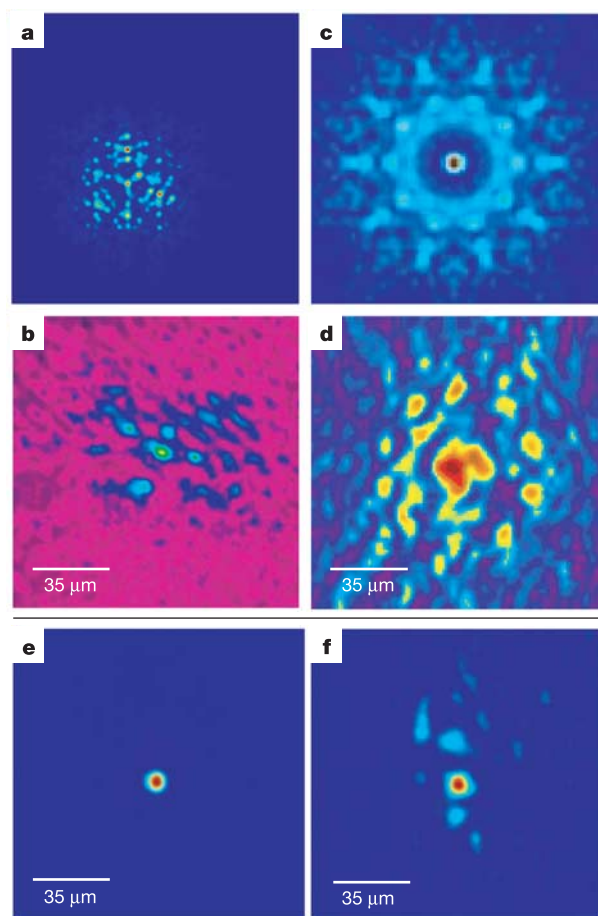


Figure 2 | Discrete diffraction and lattice solitons in a photonic quasicrystal. **a–d**, Linear tunnelling (‘discrete diffraction’) in a decagonal photonic quasicrystal: **a**, **c**, simulation results of the probe beam exiting the photonic quasicrystal for an input beam launched at the two different sites marked by yellow arrows in Fig. 1a; **b**, **d**, the corresponding experimental results (note the nicely resolved ring of 10 spots in **d**). **e**, **f**, Formation of a soliton in a quasicrystal when the intensity of the probe beam (that linearly diffracted in **d**) is sufficiently increased (with the lattice conditions unchanged). **e**, Input face and **f**, output face of the soliton beam.

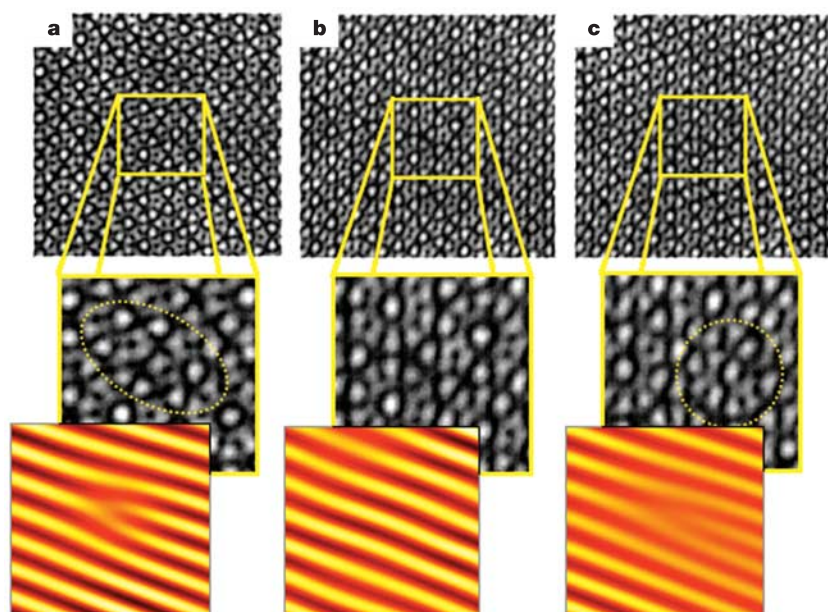


Figure 3 | Creation and healing of a dislocation in a decagonal quasicrystal. **a**, Non-interacting quasicrystal with a $(1,0,0,0)$ dislocation. **b**, With nonlinear interaction in the same crystal the dislocation disappears. The resulting ‘healed’ quasicrystal is again a ground state of the system, but is in general a ground state that differs from the original one (in Fig. 1b) by phason rearrangements throughout the whole plane (long-range effect). **c**, With increased power to the vortex beam the nonlinearity can no longer overcome the defect within the propagation distance, and the dislocation is again visible. The top panels show the actual quasicrystal. The middle panels are a magnification of the middle section of the top panels. The bottom panels show one of the filtered harmonic components clearly visualizing the dislocation.

observed²⁰. However, we find, numerically and experimentally, that an interacting decagonal photonic quasicrystal of the self-focusing type disintegrates rapidly when the waves inducing it are extraordinarily polarized. On the other hand, when the nonlinearity is of the self-defocusing type (when a negative field is applied to the photorefractive SBN:75 crystal), the interacting quasicrystal is much more stable, and the structure does not disintegrate. We pursue this avenue and, for the interacting lattice experiments, we induce a photonic quasicrystal exhibiting the photorefractive self-defocusing nonlinearity^{26–28}. Repeating the Brillouin zone spectroscopy and Bragg diffraction experiments reveals pictures very similar to Fig. 1b and d, thus ensuring that we still have a fully-fledged photonic quasicrystal when the lattice is dynamically interacting.

Having created a stable interacting photonic quasicrystal, we can use it to study the dynamics of defects²⁹ by observing the motion of dislocations and the relaxation of phason strain, as these propagate in the quasicrystal. We induce defects in the photonic quasicrystal by changing one of the plane-wave writing beams into a 2π vortex beam. This new combination of writing beams results in an intensity

pattern, and therefore a photonic quasicrystal, with a dislocation with Burgers vector $(1,0,0,0)$. When the quasicrystal is formed by non-interacting ordinarily polarized waves, the dislocation is visible even when the vortex beam is very weak and even after a considerable propagation distance (5 mm in our experiment; Fig. 3a). This is due to the fact that, without nonlinear interaction, there is no dynamics and the photonic crystal remains fixed and cannot ‘heal’ itself. However, when the crystal is induced by interacting extraordinarily polarized waves, ‘atoms’ (crystal sites) can change their positions and even merge or split, so the quasicrystal can rearrange itself—through ‘atomic’ diffusion and local configurational rearrangements—back into the lower-energy structure of a perfect defect-free quasicrystal (Fig. 3b). Finally, when the vortex beam inducing the dislocation is of high intensity, the nonlinear interaction between the sites of the quasicrystal can no longer heal the dislocation within the propagation distance in the crystal, and the dislocation remains all the way to the output (Fig. 3c).

At this point, it is essential to identify whether or not the defect indeed completely disappears, and to find its exact location. In a

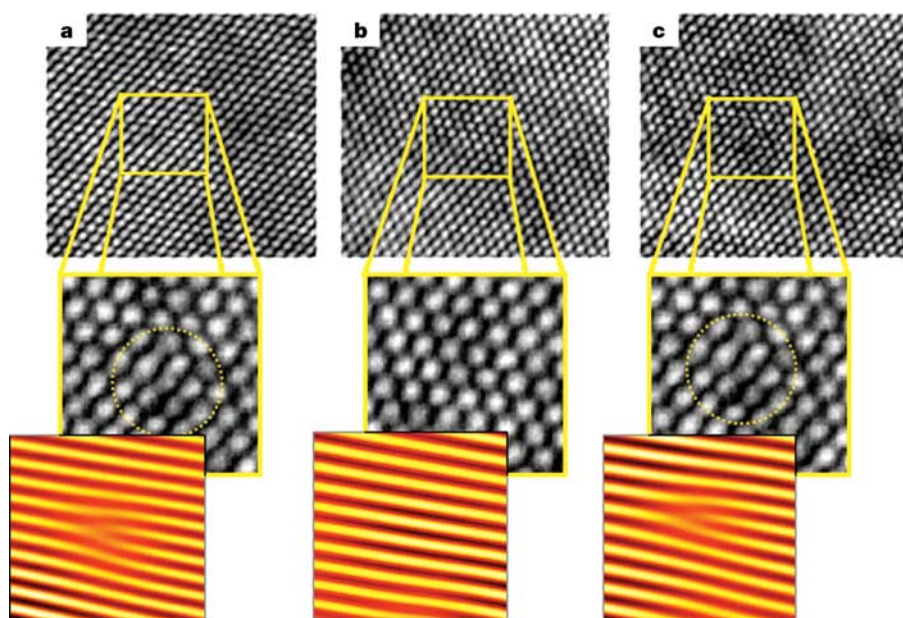


Figure 4 | Creation and healing of a dislocation in a periodic hexagonal crystal. Panels **a–c** are as for Fig. 3. **a**, The non-interacting periodic crystal with a $(1,0)$ dislocation. **b**, Note that as in the quasiperiodic case (Fig. 3), the resulting ‘healed’ crystal is in a different ground state, but here ground states can differ only by a simple shift of the pattern. **c**, As for Fig. 3c.

periodic crystal, it is rather easy to identify a dislocation, but in a quasicrystal it is not as apparent. We therefore identify the dislocations by filtering pairs of Bragg peaks in the Fourier transform image of the quasicrystal, one at a time, and then perform an inverse Fourier transform to visualize the dislocations present in each individual wave²⁹. The results, depicted in the bottom panels of Fig. 3, show clearly the defect and its location in Fig. 3a and c, and confirm that the dislocation has fully disappeared in Fig. 3b. Finally, for illustrating the difference from periodic crystals, we use three writing beams to create a hexagonal crystal with an embedded dislocation with Burgers vector (1,0), and repeat the defect dynamics experiments under similar nonlinear conditions (Fig. 4).

We recall that far away from a dislocation, both in the periodic crystal (Fig. 4a) and in the quasicrystal (Fig. 3a), the pattern is in a minimum-energy ground state, but it is in a different ground state in every direction away from the dislocation. In a periodic crystal, ground states can differ only by a shift of the crystal (a non-lattice translation). When going around the dislocation, the crystal is gradually shifted, and upon completion of a full circle and a return to the starting point, one finds that the crystal has shifted by an integer number of lattice translations (a Burgers vector). In a quasicrystal, there is greater freedom. Ground states can differ by more than simple shifts; they can also differ by rearrangements of the relative positions of the atoms (or 'tile flips' had the crystal been decorated with tiles). These are the so-called phason flips ('phasons')—a terminology that reflects the fact they arise from changing the relative phases, under certain constraints, of the different waves making up the crystal. In Fig. 3b, we clearly observe the long range consequence of phason flips that are induced in different directions around the dislocation. Particularly interesting is the predicted fact that, when a dislocation glides or climbs away, the quasicrystal cannot be as easily patched up as the periodic crystal, and a phason trail of local atomic rearrangements is left behind, which must then 'heal' through a sequence of phason flips until all 'atoms' are back in their proper positions.

The resulting 'healed' quasicrystal is again a decagonal ground state of the system, as can readily be observed in Fig. 3b, but is in general a ground state that differs from the original one by phason rearrangements throughout the whole plane (a 'long-range' effect). The comparison with the periodic hexagonal lattice of Fig. 4 is remarkable: in the periodic case, half a line of 'atoms' can simply be shifted to replace the line next to it; thus, as the defect in the hexagonal lattice of Fig. 4 moves, the crystal rearranges locally (through the so-called Peierls mechanism) and no long-range memory of the defect is left behind. Our results, depicted in Fig. 3, clearly demonstrate experimentally the creation and healing of a dislocation in a quasicrystal, and the consequence of phason flips, under controlled and engineered conditions—a 20-year-old theoretical prediction^{2,3}.

This work paves the way for a variety of experiments on linear and nonlinear wave dynamics in quasicrystals: these include direct imaging of the fractal band structure of quasicrystals, linear and nonlinear localization experiments, experiments with gap solitons, vortex dynamics in quasicrystals, modulation instability, and wave collapse in a quasiperiodic crystal. Especially interesting is the possibility of studying the dynamics of multiple defect states in interacting nonlinear optical quasicrystals and the relaxation of phason strain by observing individual phason flips. Finally, as in three-dimensional photonic crystals³⁰, we envision a three-dimensional realization of our current experiments as a means to study the details of wave dynamics in photonic structures that are quasiperiodic in all three dimensions.

Received 14 November 2005; accepted 9 March 2006.

1. Shechtman, D., Blech, I., Gratias, D. & Cahn, J. W. Metallic phase with long-range orientational order and no translational symmetry. *Phys. Rev. Lett.* **53**, 1951–1953 (1984).

2. Levine, D., Lubensky, T. C., Ostlund, S., Ramaswamy, S. & Steinhardt, P. J. Elasticity and dislocations in pentagonal and icosahedral quasicrystals. *Phys. Rev. Lett.* **54**, 1520–1523 (1985).
3. Socolar, J. E. S., Lubensky, T. C. & Steinhardt, P. J. Phonons, phasons, and dislocations in quasicrystals. *Phys. Rev. B* **34**, 3345–3360 (1986).
4. Rotenberg, E., Theis, W., Horn, K. & Gille, P. Quasicrystalline valence bands in decagonal AlNiCo. *Nature* **406**, 602–605 (2000).
5. Kaliteevski, M. A. *et al.* Two-dimensional Penrose-tiled photonic quasicrystals: from diffraction pattern to band structure. *Nanotechnology* **11**, 274–280 (2000).
6. Dal Negro, L. *et al.* Light transport through the band-edge states of Fibonacci quasicrystals. *Phys. Rev. Lett.* **90**, 055501 (2003).
7. Chan, Y. S., Chan, C. T. & Liu, Z. Y. Photonic band gaps in two dimensional photonic quasicrystals. *Phys. Rev. Lett.* **80**, 956–959 (1998).
8. Edagawa, K. & Suzuki, K. High resolution transmission electron microscopy observation of thermally fluctuating phasons in decagonal Al-Cu-Co. *Phys. Rev. Lett.* **85**, 1674–1677 (2000).
9. Efremidis, N. K., Sears, S., Christodoulides, D. N., Fleischer, J. W. & Segev, M. Discrete solitons in photorefractive optically induced photonic lattices. *Phys. Rev. E* **66**, 046602 (2002).
10. Fleischer, J. W., Carmon, T., Segev, M., Efremidis, N. K. & Christodoulides, D. N. Observation of discrete solitons in optically induced real time waveguide arrays. *Phys. Rev. Lett.* **90**, 023902 (2003).
11. Fleischer, J. W., Segev, M., Efremidis, N. K. & Christodoulides, D. N. Observation of two-dimensional discrete solitons in optically induced nonlinear photonic lattices. *Nature* **422**, 147–150 (2003).
12. Guidoni, L., Triché, C., Verkerk, P. & Grynberg, G. Quasiperiodic optical lattices. *Phys. Rev. Lett.* **79**, 3363–3366 (1997).
13. Lifshitz, R. & Petrich, D. M. Theoretical model for Faraday waves with multiple-frequency forcing. *Phys. Rev. Lett.* **79**, 1261–1264 (1997).
14. Zeng, X. *et al.* Supramolecular dendritic liquid quasicrystals. *Nature* **428**, 157–160 (2004).
15. Clausen, C. B., Kivshar, Y. S., Bang, O. & Christiansen, P. L. Quasiperiodic envelope solitons. *Phys. Rev. Lett.* **83**, 4740–4743 (1999).
16. Man, W., Megens, M., Steinhardt, P. J. & Chaikin, P. M. Experimental measurement of the photonic properties of icosahedral quasicrystals. *Nature* **436**, 993–996 (2005).
17. Bratfalean, R. T., Peacock, A. C., Broderick, N. G. R. & Gallo, K. Harmonic generation in a two-dimensional nonlinear quasicrystal. *Opt. Lett.* **30**, 424–426 (2005).
18. Lifshitz, R., Arie, A. & Bahabad, A. Photonic quasicrystals for nonlinear optical frequency conversion. *Phys. Rev. Lett.* **95**, 133901 (2005).
19. Zoorob, M. E., Charlton, M. D. B., Parker, G. J., Baumberg, J. J. & Netti, M. C. Complete photonic bandgaps in 12-fold symmetric quasicrystals. *Nature* **404**, 740–743 (2000).
20. Martin, H., Eugenieva, E. D., Chen, Z. & Christodoulides, D. N. Discrete solitons and soliton-induced dislocations in partially coherent photonic lattices. *Phys. Rev. Lett.* **92**, 123902 (2004).
21. Bartal, G. *et al.* Brillouin-zone spectroscopy of nonlinear photonic lattices. *Phys. Rev. Lett.* **94**, 163902 (2005).
22. Russell, P. St J. Optics of Floquet-Bloch waves in dielectric gratings. *Appl. Phys. B* **39**, 231–246 (1986).
23. Christodoulides, D. N. & Joseph, R. I. Discrete self-focusing in nonlinear arrays of coupled waveguides. *Opt. Lett.* **13**, 794–796 (1988).
24. Christodoulides, D. N., Lederer, F. & Silberberg, Y. Discretizing light behaviour in linear and nonlinear waveguide lattices. *Nature* **424**, 817–823 (2003).
25. Eisenberg, H. S., Silberberg, Y., Morandotti, R., Boyd, A. R. & Aitchison, J. S. Discrete spatial optical solitons in waveguide arrays. *Phys. Rev. Lett.* **81**, 3383–3386 (1998).
26. Segev, M., Valley, G. C., Crosignani, B., DiPorto, P. & Yariv, A. Steady-state spatial screening solitons in photorefractive materials with external applied field. *Phys. Rev. Lett.* **73**, 3211–3214 (1994).
27. Christodoulides, D. & Carvalho, M. Bright, dark, and gray spatial soliton states in photorefractive media. *J. Opt. Soc. Am. B* **12**, 1628–1633 (1995).
28. Efremidis, N. K. *et al.* Two-dimensional optical lattice solitons. *Phys. Rev. Lett.* **91**, 213906 (2003).
29. Barak, G. & Lifshitz, R. Dislocation dynamics in a dodecagonal quasiperiodic structure. *Phil. Mag.* **86**, 1059–1064 (2006).
30. Joannopoulos, J. D., Meade, R. D. & Winn, J. N. *Photonic Crystals: Molding the Flow of Light* (Princeton, Princeton Univ. Press, 1995).

Acknowledgements This work was supported by the Israeli Science Foundation, the Israel-USA Binational Science Foundation and the German-Israeli DIP Project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S. (msegev@technion.ac.il).

LETTERS

Electron–phonon coupling reflecting dynamic charge inhomogeneity in copper oxide superconductors

D. Reznik^{1,2}, L. Pintschovius¹, M. Ito³, S. Iikubo³, M. Sato³, H. Goka⁴, M. Fujita⁴, K. Yamada⁴, G. D. Gu⁵ & J. M. Tranquada⁵

The attempt to understand copper oxide superconductors is complicated by the presence of multiple strong interactions in these systems. Many believe that antiferromagnetism is important for superconductivity, but there has been renewed interest in the possible role of electron–lattice coupling^{1–4}. The conventional superconductor MgB₂ has a very strong electron–lattice coupling, involving a particular vibrational mode (phonon) that was predicted by standard theory and confirmed quantitatively by experiment⁵. Here we present inelastic scattering measurements that show a similarly strong anomaly in the Cu–O bond-stretching phonon in the copper oxide superconductors La_{2–x}Sr_xCuO₄ (with $x = 0.07, 0.15$). Conventional theory does not predict such behaviour. The anomaly is strongest in La_{1.875}Ba_{0.125}CuO₄ and La_{1.48}Nd_{0.4}Sr_{0.12}CuO₄, compounds that exhibit spatially modulated charge and magnetic order, often called stripe order⁶; it occurs at a wave vector corresponding to the charge order. These results suggest that this giant electron–phonon anomaly, which is absent in undoped and over-doped non-superconductors, is associated with charge inhomogeneity. It follows that electron–phonon coupling may be important to our understanding of superconductivity, although its contribution is likely to be indirect.

Ordinary metals can be understood within the conventional Fermi-liquid picture, where electron-like quasiparticles populate bands in energy-momentum space up to the cut-off at the Fermi energy. The dispersion of these bands is determined by the crystal structure and chemistry. Lattice vibrations couple to electrons because displacements of atoms from their equilibrium positions alter the band dispersions, lowering or raising the total electronic energy. Calculations based on this standard model have been very successful in explaining many properties associated with electron–phonon coupling in conventional metals, including phonon-mediated superconductivity. In the case of copper oxide superconductors, however, such electron–phonon interactions appear to be too weak to explain the high superconducting transition temperatures (T_c). Furthermore, the applicability of the Fermi-liquid model to the copper oxides has been questioned.

In one alternative approach, it has been argued that the competition between the kinetic energy of charge carriers and the antiferromagnetic superexchange between magnetic moments on neighbouring Cu atoms leads to charge segregation^{7–9}. Experimental support for such a view was provided when static charge segregation in the form of stripes was first discovered in La_{1.48}Nd_{0.4}Sr_{0.12}CuO₄ (ref. 6), where a small structural modification provides a pinning potential for the stripes. The fact that static stripes compete with superconductivity has led many to question the relevance of charge inhomogeneity to understanding the superconductivity.

The possible existence of dynamic charge inhomogeneity has been difficult to establish in the absence of a clear experimental signature⁹.

Charge inhomogeneity with a periodic modulation should affect interatomic Coulomb forces and soften lattice vibrations at the same wave vector. For example, in compounds with stripes one expects naively that the strongest effect should be in the longitudinal bond-stretching mode whose polarization pattern matches the lattice deformation induced by the charge inhomogeneity (Fig. 1). This idea has motivated us to test whether some feature of the relevant phonon mode might provide a useful measure of dynamic charge inhomogeneity. Although there have been previous reports of phonon anomalies in copper oxide superconductors^{10–12}, and unexpected phonon dispersions have been observed in La_{1.69}Sr_{0.31}NiO₄ with diagonal stripe order¹³, no one has made the crucial test of comparing electron–phonon coupling in copper oxide compounds at doping levels associated variously with static stripe order, superconductivity and Fermi-liquid behaviour.

Thus, we began our study by investigating the bond-stretching vibration along the Cu–O bond direction in La_{1.48}Nd_{0.4}Sr_{0.12}CuO₄ (ref. 14) and La_{1.875}Ba_{0.125}CuO₄ (ref. 15), compounds in which diffraction measurements have demonstrated the presence of static stripes. In addition, a soft-X-ray resonant-diffraction study¹⁶ of La_{1.875}Ba_{0.125}CuO₄ has recently provided direct evidence for concomitant charge order, below $T_s \approx 60$ K, with the modulation wave vector $\mathbf{q}_{co} = (0.25\ 0\ 0)$. We performed most of our measurements

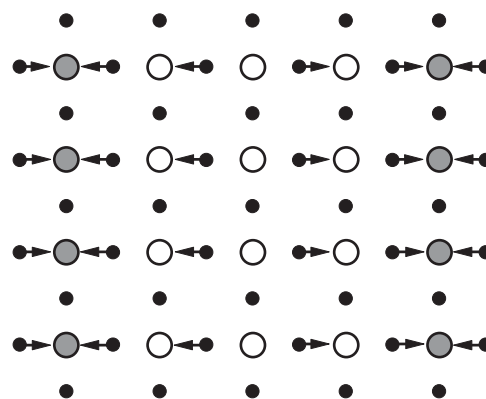


Figure 1 | Displacement pattern of the oxygen ions for the phonon with $\mathbf{q} = (0.25\ 0\ 0)$ propagating perpendicular to the stripes. Arrows represent atomic displacements. Large circles correspond to copper and small ones to oxygen. Open large circles represent hole-poor antiferromagnetic regions, while the filled circles represent the hole-rich lines.

¹Forschungszentrum Karlsruhe, Institut für Festkörperphysik, POB 3640, D-76021 Karlsruhe, Germany. ²Laboratoire Léon Brillouin, CE Saclay, F-91191 Gif-sur-Yvette Cedex, France. ³Department of Physics, Division of Materials Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602, Japan. ⁴Institute for Material Research, Tohoku University, Katahira, Aoba-ku, Sendai 980-8577, Japan. ⁵Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, New York 11973-5000, USA.

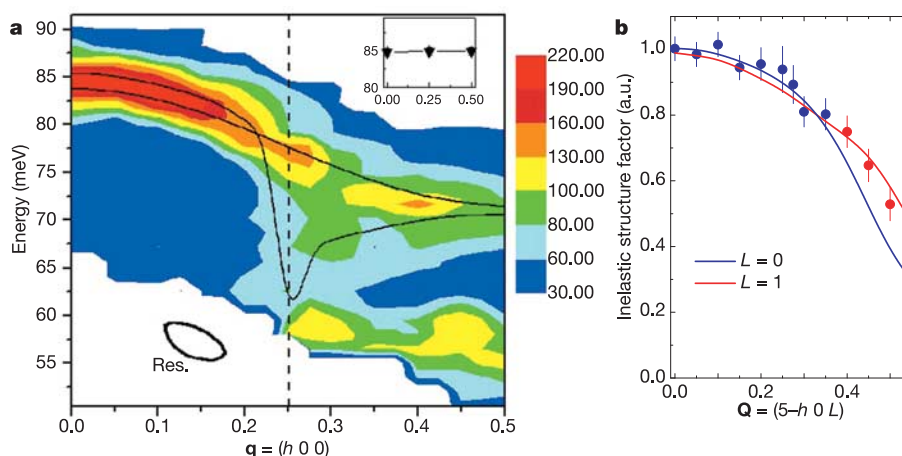


Figure 2 | Bond-stretching phonon branch in $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$. The experiments were performed on the 1T triple-axis spectrometer at the ORPHEE reactor at LLB, Saclay. A Cu220 reflection was used as the monochromator in order to achieve high resolution, while the 002 pyrolytic graphite analyser reflection was fixed at 14.7 meV. The wave vectors are given in reciprocal lattice units of $(2\pi/a \ 2\pi/b \ 2\pi/c)$, where $a = b = 3.78 \text{ \AA}$, and $c = 13.18 \text{ \AA}$. The scattering cross-section was measured in the Brillouin zone near the reciprocal lattice vector $\mathbf{G} = (5 \ 0 \ L)$ (where $L = 0, 1$ or 2) to maximize accessible dynamic structure factor of the bond-stretching phonons. For neutron momentum transfer $\hbar\mathbf{Q}$, $\mathbf{q} = \mathbf{G} - \mathbf{Q}$. The single crystal sample of $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ (ref. 15) weighing 5 g was mounted on the spectrometer with the b axis vertical. For temperature control, each sample was mounted in a closed-cycle refrigerator. **a**, Colour-coded contour plot of the intensities in arbitrary units observed on $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ at $T = 10 \text{ K}$.

at low temperature, $T \approx 10 \text{ K}$, in the ordered stripe phase.

Figure 2a shows a colour-coded map of the bond-stretching phonon spectra in $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ at $T = 10 \text{ K}$ for phonon wave vectors $\mathbf{q} = (h \ 0 \ 0)$. It is immediately clear that the well-defined branch dispersing downward from the zone centre acquires a broad low-energy tail near to $\mathbf{q}_{\text{co}}$. In that vicinity, the individual scans can be fitted with a two-peak structure (see Fig. 3a, b and discussion below), with the line shape returning to a single peak at the zone boundary. The integrated intensity of the entire feature agrees very well with the calculated¹⁷ structure factor of the bond-stretching mode, as shown in Fig. 2b.

Figure 3 compares individual scans through the phonon mode at $\mathbf{q} = (0.15 \ 0 \ 0)$ and at $\mathbf{q}_{\text{co}}$; these points are separated by slightly more than the full-width at half-maximum (FWHM) of the instrumental $\mathbf{Q}$ -resolution. Whereas at $\mathbf{q} = (0.15 \ 0 \ 0)$ there is a single sharp peak, at $\mathbf{q}_{\text{co}}$ the main peak is much weaker and the missing intensity appears in the low energy tail. These latter spectra can be fitted with two displaced peaks with different energy widths but comparable integrated intensities. Extending the two-peak analysis to the full $\mathbf{q}$ range (and assuming resolution-limited energy widths at small $\mathbf{q}$) leads to a model of two phonon branches (instead of one), as indicated by the solid lines in Fig. 2a. One branch has a monotonic, cosine-like downward dispersion in the $[100]$ direction. The other branch is highly anomalous in that it drops abruptly by more than 10 meV at about $\mathbf{q}_{\text{co}}$ and then rises back up towards the zone boundary to merge with the other branch; it also has an anomalously large energy width. The inferred drop is much sharper than the instrumental resolution in the longitudinal direction, and also in the transverse direction (not shown), so that the average over wave vectors within the resolution ellipsoid produces the broad low-energy tail at $\mathbf{q}_{\text{co}}$. A splitting of the bond-stretching mode is compatible with the crystal symmetry in the ordered stripe phase. The inferred variation between the two branches could indicate a substantial difference for phonons propagating parallel and perpendicular to the stripes. (The symmetry of the crystal also ensures that we would average over both components.)

The intensities above and below 60 meV are associated with plane-polarized Cu–O bond-stretching vibrations and bond-bending vibrations, respectively. Black lines are dispersion curves evaluated from two-peak fits to the data (one of them has a cosine behaviour). The white area at the lower left corner of the diagram was not accessible in this experiment. The ellipse illustrates the instrumental resolution ('Res.'). The inset shows the dispersion in the $[110]$ direction (same x and y axes as main panel). The dashed line represents the charge-ordering wave vector. **b**, The dots denote inelastic structure factors of the bond-stretching modes as deduced from the integrated intensities observed for wave vectors $\mathbf{Q} = (5 - h \ 0 \ 0)$ (blue) and $\mathbf{Q} = (5 - h \ 0 \ 1)$ (red); error bars represent ± 1 s.d. The data were normalized to the value at $\mathbf{Q} = (5 \ 0 \ 0)$. The solid lines were calculated from a shell model¹⁷. We note that the inelastic structure factors depend very little on the value of L except for wave vectors close to the zone boundary. a.u., arbitrary units.

Consistent with the model of stripes aligned along Cu–O bonds, the dispersion along the $[110]$ direction is much more normal. The inset of Fig. 2a shows that the bond-stretching branch in the $[110]$ direction is completely flat. It has a broad linewidth, which may reflect bond stiffness variations due to the inhomogeneous charge distribution. Also, the anomalous lineshape along $[100]$ is independent of the out-of-plane $[001]$ component of $\mathbf{q}$, in agreement with the stripe dynamics being uncorrelated between neighbouring CuO_2 planes. Furthermore, the softening and broadening effects are strongest in the bond-stretching branch: detailed measurements did not find any anomalies in the bond-bending branch (appearing near 60 meV in Fig. 2a).

Temperature dependence provides further confirmation of the unusual nature of the phonon line shape near $\mathbf{q}_{\text{co}}$. The anomaly is strongest at the lowest temperature, 10 K (Fig. 3a, b). At 330 K (Fig. 3c), the low-energy tail is greatly reduced and the peak is enhanced, though even at this temperature the conventional line shape, as indicated by the curve at $\mathbf{q} = (0.15 \ 0 \ 0)$, is not fully recovered. The conventional anharmonic behaviour of phonons is to broaden and soften with increasing temperature, not sharpen as we observe.

To test the doping dependence of the anomaly, we have measured large single crystal samples of underdoped ($T_c = 20 \text{ K}$), optimally doped ($T_c = 35 \text{ K}$) and overdoped non-superconducting $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ ($x = 0.07, 0.15$ and 0.3 , respectively). Figure 4a compares low-temperature scans at $\mathbf{q} = (0.25 \ 0 \ 0)$ for all of these samples, together with a previous result for La_2CuO_4 . As one can see, the peaks are considerably broader for $x = 0.07$ and 0.15 than for the undoped and over-doped samples. In Fig. 4b we compare the dispersion of the bond-stretching branch for all compounds measured in the current study with published data on $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ ($T_c = 66 \text{ K}$)¹⁸ and $\text{YBa}_2\text{Cu}_3\text{O}_{6.95}$ ($T_c = 93 \text{ K}$)¹⁹. The superconducting and stripe phase compounds are characterized by substantial deviations from the simple cosine dispersion, with the biggest effects at $\mathbf{q} = (0.25 - 0.3 \ 0 \ 0)$ (Fig. 4c), and with narrow linewidths found for the overdoped non-superconducting $\text{La}_{1.7}\text{Sr}_{0.3}\text{CuO}_4$. We note that, consistent with the trend found here, a similar difference from cosine-like

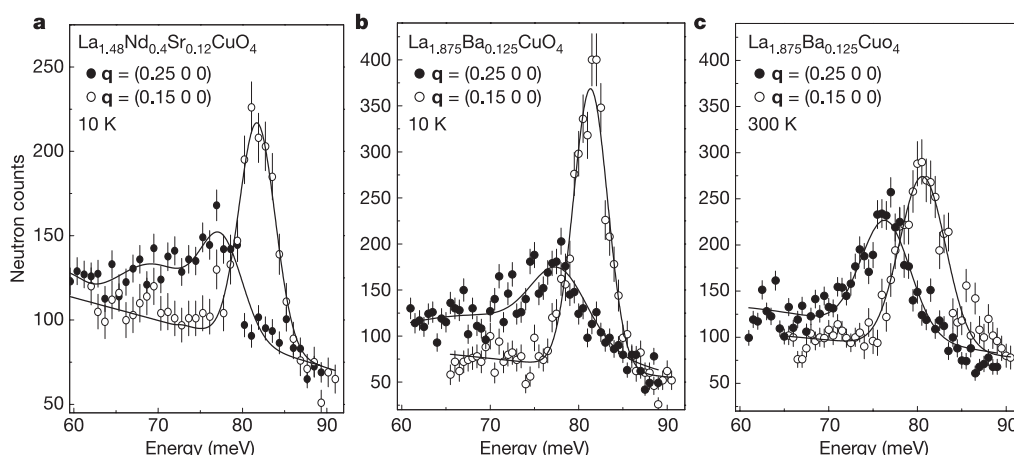


Figure 3 | Representative energy scans at 10 K and 330 K. **a–c**, Energy scans taken on $\text{La}_{1.48}\text{Nd}_{0.4}\text{Sr}_{0.12}\text{CuO}_4$ at $T = 10$ K (**a**) and on $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ (**b**, **c**) at 10 K (**a**, **b**) and 330 K (**c**). Error bars represent ± 1 s.d. The sample of $\text{La}_{1.48}\text{Nd}_{0.4}\text{Sr}_{0.12}\text{CuO}_4$ (ref. 14) consisted of six single crystals coaligned with the c -axis vertical. The phonon at $\mathbf{q} = (0.15\ 0\ 0)$ is ‘normal’, that is, it has a gaussian lineshape on top of a linear background, which results from multiphonon and incoherent scattering and has no

strong dependence on $\mathbf{q}$. Its intensity reduction from 10 K (**b**) to 330 K (**c**) is consistent with the Debye-Waller factor. At $\mathbf{q} = (0.25\ 0\ 0)$, there is extra intensity on top of the background in the tail of the main peak. It results from one-phonon scattering that extends to the lowest investigated energies, while the peak intensity is greatly suppressed as discussed in the text. The effect is reduced but does not disappear on heating from 10 K to 330 K. Note that $\mathbf{G} = (5\ 0\ 2)$ for **a** and $(5\ 0\ 0)$ for **b** and **c**.

dispersion, together with a large energy broadening, has been reported for superconducting $\text{HgBa}_2\text{CuO}_{4+\delta}$ (ref. 20) at the same wave vectors. Thus, there seems to be a general correlation between the occurrence of strong electron–phonon coupling effects in the bond-stretching mode half-way to the zone boundary and a hole concentration in the range compatible with superconductivity in the copper oxides.

In terms of temperature dependence, the onset of the anomaly moves to lower temperatures with increasing x in both $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ and $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$. For example, the anomaly is only slightly reduced in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ when the temperature is raised to 300 K (ref. 18), but in $\text{YBa}_2\text{Cu}_3\text{O}_{6.95}$ the normal dispersion is already recovered by 200 K (refs 19, 21). The dispersions of the same phonon branch in the non-superconducting undoped La_2CuO_4 and overdoped $\text{La}_{1.7}\text{Sr}_{0.3}\text{CuO}_4$ do not show any sign of the anomaly at any temperature.

We believe that McQueeney *et al.*¹⁰ and Chung *et al.*¹¹ reported

manifestations of the same effect in $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_4$ and $\text{YBa}_2\text{Cu}_3\text{O}_{6.95}$, respectively, that are compatible with our observations; however, our measurements, performed with much higher resolution and better signal-to-background ratio, do not confirm their evidence for unit cell doubling. Compared with ref. 22, the anomaly appears to be stronger in the present study possibly owing to higher sample quality. Other previous reports of phonon anomalies in the copper oxides focused on an entirely different effect—softening of the so-called ‘half breathing’ mode at the zone boundary, $\mathbf{q} = (0.5\ 0\ 0)$, with increased doping¹². This effect is distinct from the one we report here as it continues into the overdoped regime²³; in fact it is responsible for the downward cosine dispersion that we consider as ‘normal’ behaviour, as it is predicted by conventional calculations³.

In the case of superconducting MgB_2 , one of the highest- T_c superconductors ($T_c = 39$ K) for which electron–phonon coupling

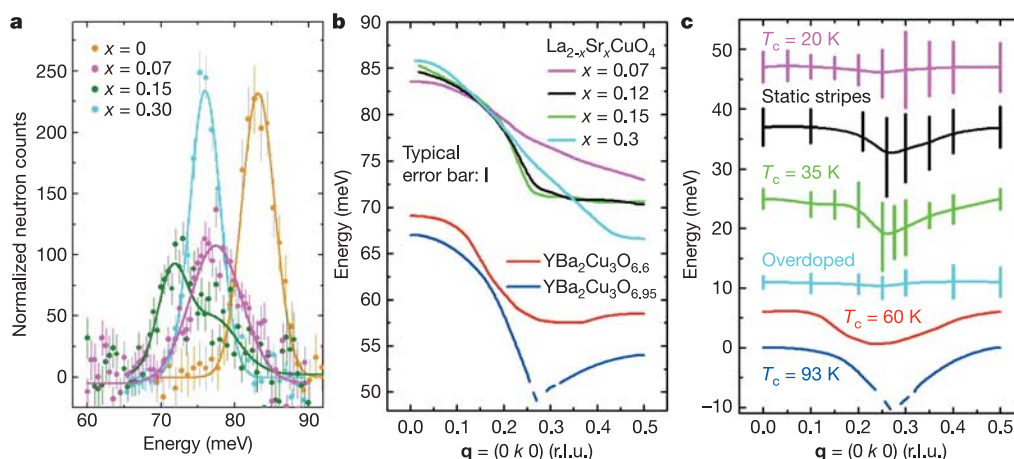


Figure 4 | Correlation of the phonon anomaly with stripe order and superconductivity. **a**, Comparison of energy scans taken at 10 K on $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ at $\mathbf{q} = (0.25\ 0\ 0)$, after subtraction of linear background. Solid lines are guides to the eye; error bars represent ± 1 s.d. The resolution was somewhat lower in the measurement on the $x = 0$ sample (L.P. W. Reichardt and A. Yu Rumiantsev, unpublished results), where Cu111, rather than Cu220, monochromator was used. **b**, Dispersion of the anomalous branch in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ ($x = 0.07, 0.12$ and 0.15) and the [010] bond-stretching branch along the chain direction in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$ ($T_c = 66$ K)¹⁸ and $\text{YBa}_2\text{Cu}_3\text{O}_{6.95}$ ($T_c = 93$ K)¹⁹. The $x = 0.12$ compound was $\text{La}_{1.48}\text{Nd}_{0.4}\text{Sr}_{0.12}\text{CuO}_4$. The curve for $\text{YBa}_2\text{Cu}_3\text{O}_{6.95}$ is only approximate,

as the anomaly is so strong near $\mathbf{q} = (0\ 0.25\ 0)$ that there is no well defined peak there (see Fig. 1 in ref. 19). The difference in energies at $\mathbf{q} = 0$ results from slightly different Cu–O bond lengths. Dashed blue line represents the region where the bond-stretching mode mixes with others and its precise energy cannot be established²¹. **c**, Difference between the downward cosine dispersion, fitted to the zone centre and the zone boundary for each sample, and the dispersions in **b**; vertical lines represent peak widths. For the $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ data the width information is not available because the data were taken in the defocusing scattering geometry with relatively poor energy resolution. Curves are offset for clarity. Dashed blue line is the same as in **b**.

is agreed to be responsible for electron pairing, very large (10–15 meV FWHM) phonon linewidths have been measured⁵. Both the linewidths and the dispersions are in good agreement with calculations based on the conventional theory of electronic structure (using the local-density approximation). The effect there is explained in terms of a Kohn anomaly.

The sharp dispersion and strong line broadening that we observe in the bond-stretching mode of superconducting copper oxides also looks similar to a Kohn anomaly or charge-density-wave instability. So can our results be understood within conventional models that assume electronic homogeneity? The answer seems to be no. Several independent electron–phonon coupling calculations fail to predict the huge effect that is observed here^{2–4}. They all give a very similar Fermi surface and adequately account for many non-trivial features of the phonon spectrum, such as the downward dispersion of the normal bond-stretching branch at optimal doping and the modest electron–phonon coupling for the bond-buckling branch dispersing from the B_{1g} Raman mode near 42 meV; however, these calculations greatly underestimate the renormalization of the anomalous branch. For example, refs 2 and 3 give an electron–phonon coupling constant for the bond-stretching branch that is much smaller than that for the bond-buckling branch, which is renormalized by at most 1 meV (ref. 24).

Of course, strong electron–electron correlations are expected to be important in copper oxides for electronic states near the Fermi level, and these effects are not properly described in the conventional theory. An example of a model intended to capture correlation effects is the so-called t - J model. Extensions of that model to include electron–phonon coupling allow a reasonable description of the doping dependence of the half-breathing mode²⁵. We note that Ishihara and Nagaosa²⁶ were able to get a substantial dip in the bond-stretching branch at $\mathbf{q} \approx (0.3 \ 0 \ 0)$; however, they had to use an off-diagonal electron–phonon coupling strength an order of magnitude larger than that estimated in ref. 25.

The empirical result found here is that the largest dispersion dip and line broadening occur in samples that exhibit static stripe order. Theoretically, Kaneshita *et al.*²⁷ have predicted that such behaviour should occur at $\mathbf{q}_{co}$ for the bond-stretching mode propagating in a direction perpendicular to the stripes. Of course, the phonons that we have studied are at fairly high energy, and the response at such energies need not show great sensitivity to the presence or absence of correlations at zero energy (that is, static order). In the superconducting copper oxides, we believe that the anomaly reflects a response to dynamic charge inhomogeneity. The orientation of the response with respect to the charge modulation wave vector is not clear-cut. Also, on the basis of the stripe picture the strongest anomaly in $\text{La}_{1.93}\text{Sr}_{0.07}\text{CuO}_4$ should be at $h = 0.14$ (magnetic incommensurability is 0.07; ref. 28), and not at 0.3 as observed in the experiment (Fig. 4c). Furthermore, a stripe-based interpretation of the anisotropic magnetic response²⁹ measured in $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ would suggest dynamic stripes oriented along the [010] direction, parallel to the Cu–O chains, and this is the same direction in which the large phonon anomaly appears. Thus, the possible role of Fermi-surface nesting effects associated with electronic states with wave vectors parallel to charge stripes should be considered³⁰.

To conclude, our experiments show strong bond-stretching phonon anomalies common to stripe-ordered and superconducting copper oxides. These results imply the existence of dynamic charge inhomogeneity due to strong electronic correlations, as the observed effects are absent in calculations using conventional techniques. The disappearance of the phonon anomaly at the non-superconducting extremes of doping suggests that electron–phonon coupling may be important to the mechanism of superconductivity, although the contribution may be indirect.

Received 16 November 2005; accepted 2 March 2006.

1. Lanzara, A. *et al.* Evidence for ubiquitous strong electron-phonon coupling in

- high-temperature superconductors. *Nature* **412**, 510–514 (2001).
2. Devereaux, T. P., Cuk, T., Shen, Z.-X. & Nagaosa, N. Anisotropic electron-phonon interaction in the cuprates. *Phys. Rev. Lett.* **93**, 117004 (2004).
3. Bohnen, K.-P., Heid, R. & Krauss, M. Phonon dispersion and electron-phonon interaction for $\text{YBa}_2\text{Cu}_3\text{O}_7$ from first-principles calculations. *Europhys. Lett.* **64**, 104–110 (2003).
4. Falter, C. & Hoffmann, G. A. Nonlocal electron-phonon coupling of ionic charge-fluctuation type and phonon anomalies in high-temperature superconductors. *Phys. Rev. B* **64**, 054516 (2001).
5. Baron, A. Q. R. *et al.* Kohn anomaly in MgB_2 by inelastic X-ray scattering. *Phys. Rev. Lett.* **92**, 197004 (2004).
6. Tranquada, J. M. *et al.* Evidence for stripe correlations of spins and holes in copper oxide superconductors. *Nature* **375**, 561–563 (1995).
7. Zaanen, J. & Gunnarsson, O. Charged magnetic domain lines and the magnetism of high- T_c oxides. *Phys. Rev. B* **40**, 7391–7394 (1989).
8. Machida, K. Magnetism in La_2CuO_4 based compounds. *Physica C* **158**, 192–196 (1989).
9. Kivelson, S. A. *et al.* How to detect fluctuating stripes in the high-temperature superconductors. *Rev. Mod. Phys.* **75**, 1201–1241 (2003).
10. McQueeney, R. J. *et al.* Anomalous dispersion of LO phonons in $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_4$ at low temperatures. *Phys. Rev. Lett.* **82**, 628–631 (1999).
11. Chung, J.-H. *et al.* In-plane anisotropy and temperature dependence of oxygen phonon modes in $\text{YBa}_2\text{Cu}_3\text{O}_{6.95}$. *Phys. Rev. B* **67**, 014517 (2003).
12. Pintschovius, L. *et al.* Lattice dynamical studies of HTSC materials. *Physica C* **185–189**, 156–161 (1991).
13. Tranquada, J. M., Nakajima, K., Braden, M., Pintschovius, L. & McQueeney, R. J. Bond-stretching-phonon anomalies in stripe-ordered $\text{La}_{1.69}\text{Sr}_{0.31}\text{NiO}_4$. *Phys. Rev. Lett.* **88**, 075505 (2002).
14. Ito, M. *et al.* Effects of “stripes” on the magnetic excitation spectra of $\text{La}_{1.48}\text{Nd}_{0.4}\text{Sr}_{0.12}\text{CuO}_4$. *J. Phys. Soc. Jpn* **72**, 1627–1630 (2003).
15. Fujita, M., Goka, H., Yamada, K., Tranquada, J. M. & Regnault, L. P. Stripe order, depinning, and fluctuations in $\text{La}_{1.875}\text{Ba}_{0.125}\text{CuO}_4$ and $\text{La}_{1.875}\text{Ba}_{0.075}\text{Sr}_{0.050}\text{CuO}_4$. *Phys. Rev. B* **70**, 104517 (2004).
16. Abbamonte, P. *et al.* Spatially modulated ‘Mottness’ in $\text{La}_{2-x}\text{Ba}_x\text{CuO}_4$. *Nature Phys.* **1**, 155–158 (2005).
17. Chaplot, S. L., Reichardt, W., Pintschovius, L. & Pyka, N. Common interatomic potential model for the lattice dynamics of several cuprates. *Phys. Rev. B* **52**, 7230–7242 (1995).
18. Pintschovius, L., Reichardt, W., Kläser, M., Wolf, T. & von Löhnysen, H. Pronounced in-plane anisotropy of phonon anomalies in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. *Phys. Rev. Lett.* **89**, 037001 (2002).
19. Pintschovius, L. *et al.* Oxygen phonon branches in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Phys. Rev. B* **69**, 214506 (2004).
20. Uchiyama, H. *et al.* Softening of Cu–O bond stretching phonons in tetragonal $\text{HgBa}_2\text{CuO}_{4+\delta}$. *Phys. Rev. Lett.* **92**, 197005 (2004).
21. Pintschovius, L. *et al.* Evidence for dynamic charge stripes in the phonons of optimally doped YBCO. Preprint at (http://www.arxiv.org/PS_cache/cond-mat/pdf/0308/0308357.pdf) (2003).
22. Pintschovius, L. & Braden, M. Anomalous dispersion of LO phonons in $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_4$. *Phys. Rev. B* **60**, R15039 (1999).
23. Fukuda, T. *et al.* Doping dependence of softening in the bond-stretching phonon mode of $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ ($0 \leq x \leq 0.29$). *Phys. Rev. B* **71**, 060501(R) (2005).
24. Reznik, D., Keimer, B., Dogan, F. & Aksay, I. A. q Dependence of self-energy effects of the plane oxygen vibration in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Phys. Rev. Lett.* **75**, 2396–2399 (1995).
25. Röscher, O. & Gunnarsson, O. Electron-phonon interaction in the t - J model. *Phys. Rev. Lett.* **92**, 146403 (2004).
26. Ishihara, S. & Nagaosa, N. Interplay of electron-phonon interaction and electron correlation in high-temperature superconductivity. *Phys. Rev. B* **69**, 144520 (2004).
27. Kaneshita, E., Ichioka, M. & Machida, K. Phonon anomalies due to collective stripe modes in high T_c cuprates. *Phys. Rev. Lett.* **88**, 115501 (2002).
28. Hiraka, H. *et al.* Spin fluctuations in the underdoped high- T_c cuprate $\text{La}_{1.93}\text{Sr}_{0.07}\text{CuO}_4$. *Phys. Soc. Jpn* **70**, 853–858 (2001).
29. Hinkov, V. *et al.* Two-dimensional geometry of spin excitations in the high temperature superconductor $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$. *Nature* **430**, 650–654 (2004).
30. Zhou, X. J. *et al.* One-dimensional electronic structure and suppression of d-wave node state in $\text{La}_{1.28}\text{Nd}_{0.6}\text{Sr}_{0.12}\text{CuO}_4$. *Science* **286**, 268–272 (1999).

Acknowledgements G.D.G. and J.M.T. are supported by the Office of Science, US Department of Energy. K.Y., M.F. and M.S. are supported by grants from the MEXT of Japan. D.R. thanks S. Kivelson for comments on the first version of the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.R. (Reznik@llb.saclay.cea.fr).

LETTERS

Dislocation multi-junctions and strain hardening

Vasily V. Bulatov¹, Luke L. Hsiung¹, Meijie Tang¹, Athanasios Arsenlis¹, Maria C. Bartelt¹, Wei Cai^{1,2}, Jeff N. Florando¹, Masato Hiratani¹, Moon Rhee¹, Gregg Hommes¹, Tim G. Pierce¹ & Tomas Diaz de la Rubia¹

At the microscopic scale, the strength of a crystal derives from the motion, multiplication and interaction of distinctive line defects called dislocations. First proposed theoretically in 1934 (refs 1–3) to explain low magnitudes of crystal strength observed experimentally, the existence of dislocations was confirmed two decades later^{4,5}. Much of the research in dislocation physics has since focused on dislocation interactions and their role in strain hardening, a common phenomenon in which continued deformation increases a crystal's strength. The existing theory relates strain hardening to pair-wise dislocation reactions in which two intersecting dislocations form junctions that tie the dislocations together^{6,7}. Here we report that interactions among three dislocations result in the formation of unusual elements of dislocation network topology, termed 'multi-junctions'. We first predict the existence of multi-junctions using dislocation dynamics and atomistic simulations and then confirm their existence by transmission electron microscopy experiments in single-crystal molybdenum. In large-scale dislocation dynamics simulations, multi-junctions present very strong, nearly indestructible, obstacles to dislocation motion and furnish new sources for dislocation multiplication, thereby playing an essential role in the evolution of dislocation microstructure and strength of deforming crystals⁸. Simulation analyses conclude that multi-junctions are responsible for the strong orientation dependence of strain hardening in body-centred cubic crystals.

The amount of slip produced by a propagating dislocation is quantified by its Burgers vector $\mathbf{b}$, which is equal to one of the (typically smallest) repeat vectors of the crystal lattice. Exactly what happens when two dislocations collide depends on the lengths of two lines, their collision geometry and applied stress. Most significantly, the collision outcomes are affected by the Burgers vectors of two colliding dislocations. Given that a dislocation's energy is proportional to the square of its Burgers vector, the approximate Frank energy criterion⁹ predicts that when $(\mathbf{b}_1 + \mathbf{b}_2)^2 < \mathbf{b}_1^2 + \mathbf{b}_2^2$ or, equivalently, $\mathbf{b}_1 \cdot \mathbf{b}_2 < 0$, the two lines will attract and merge into a product dislocation—a 'junction'—with Burgers vector $\mathbf{b}_j = \mathbf{b}_1 + \mathbf{b}_2$, thereby reducing the internal energy of the system. In particular, when $\mathbf{b}_2 = -\mathbf{b}_1$, the two dislocations can annihilate completely, leaving no product.

Figure 1a and b shows a junction-forming collision of two dislocation lines in a dislocation dynamics (DD) simulation (see the Methods section and Supplementary Discussion 1 for more details). The initial configuration consists of two straight dislocation lines of equal length made to intersect at their midpoints, while their endpoints are rigidly fixed (Fig. 1a). Expressed in the units of the lattice constant, the Burgers vectors of two lines are $\mathbf{b}_1 = 1/2[\bar{1}11]$ and $\mathbf{b}_2 = 1/2[1\bar{1}1]$, typical of the body-centred cubic (b.c.c.) crystals. In the DD simulation, the elastic interaction between two lines causes them to merge into a junction dislocation with Burgers vector $\mathbf{b}_j = \mathbf{b}_1 + \mathbf{b}_2 = [001]$ (Fig. 1b, see also Supplementary Video 1).

Owing to their fixed ends, the lines merge only partially (when $\mathbf{b}_2 = -\mathbf{b}_1$ the lines will partially annihilate). Bounded at its ends by two triple nodes, the resulting junction zips along the $[111]$ direction because each of the two parent dislocations is allowed to move only on its glide plane (the plane containing both the Burgers vector and the dislocation line itself) with normal vectors $\mathbf{n}_1 = (01\bar{1})$ and $\mathbf{n}_2 = (10\bar{1})$, respectively. Because most of the interacting dislocations move on non-parallel glide planes, attractive collisions zip junctions of limited length along the intersection lines of the glide planes. The frequency and strength of such pair-wise dislocation reactions tying dislocations together are believed to control the physics of strain hardening: a common phenomenon in which continued deformation increases a crystal's strength.

The existence and the important role of dislocations junctions in strain hardening has been confirmed by numerous theoretical^{6,7} and experimental^{10,11} studies and, more recently, by DD simulations^{12,13}. Very recently, analysing previously published¹⁴ and our own new simulations, we observed the formation of complex topological connections in which more than two dislocation lines merge together. Curious about possible causes for such anomalous formations, we proceeded to investigate whether attractive reactions among three or more dislocations are possible. For b.c.c. crystals, one such candidate reaction is:

$$\frac{1}{2}[\bar{1}11] + \frac{1}{2}[1\bar{1}1] + \frac{1}{2}[11\bar{1}] = \frac{1}{2}[111] \quad (1)$$

$\mathbf{b}_1 \qquad \mathbf{b}_2 \qquad \mathbf{b}_3 \qquad \mathbf{b}_4$

Given that the elastic energy stored in a dislocation's strain field is proportional to $\|\mathbf{b}\|^2$, the Frank estimate of the energy reduction

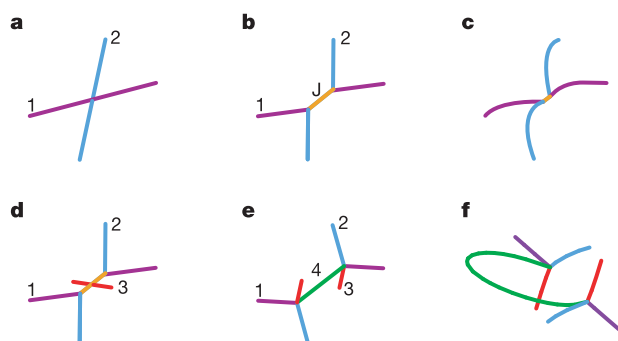


Figure 1 | Formation (zipping) and yielding of dislocation junctions in the DD simulations. **a**, Two dislocation lines are initially brought to intersection at their midpoints. **b**, Once the interaction between two lines is turned on, two lines zip a binary junction, **J**. **c**, A snapshot showing a binary junction unzipping under stress. **d**, A third line is brought to intersect the binary junction. **e**, The interaction among three lines makes them zip a long multi-junction. **f**, A snapshot showing a multi-junction acting as a Frank-Read source of dislocation multiplication.

¹Lawrence Livermore National Laboratory, University of California, Livermore, California 94550, USA. ²Department of Mechanical Engineering, Stanford University, Stanford, California 94305, USA.

in the above reaction, $(\|\mathbf{b}_1\|^2 + \|\mathbf{b}_2\|^2 + \|\mathbf{b}_3\|^2)/\|\mathbf{b}_4\|^2 = 3$, is much greater than in the binary reaction shown in Fig. 1b, $(\|\mathbf{b}_1\|^2 + \|\mathbf{b}_2\|^2)/\|\mathbf{b}_j\|^2 = 1.5$. To see whether such a reaction is indeed feasible, we now overlay on top of the binary junction a third dislocation (Fig. 1d) with Burgers vector $\mathbf{b}_3 = 1/2[111]$ and glide plane $\mathbf{n}_3 = (110)$. The DD simulation result is shown in Fig. 1e, where the third dislocation with Burgers vector $\mathbf{b}_3$ has reacted with the junction dislocation and transformed it into a $1/2[111]$ dislocation, exactly in accord with the proposed reaction (1) (Supplementary Video 2). The transformed junction segment with Burgers vector $\mathbf{b}_4$ of $1/2\langle 111 \rangle$ -type connects together three dislocations at its ends and is defined as a multi-junction. Remarkably, the resulting multi-junction extends well beyond the original length of the binary junction, corroborating our expectation that the ternary reaction (1) can result in a much greater energy reduction than the classical binary reaction.

The nodes at each end of the multi-junction that tie together four dislocation lines we term 'multi-nodes' or '4-nodes'. These 4-nodes are distinct from simple crosses of two dislocations in that all four lines entering the node have different Burgers vectors. These 4-nodes are beautifully symmetric: all four distinct Burgers vectors of $1/2\langle 111 \rangle$ -type enter the 4-node exactly once. Therefore, it is the only possible 4-node of this kind in b.c.c. crystals. Curiously, the same symmetric 4-node can be formed through four different reactions among three lines, for example $\mathbf{b}_1 + \mathbf{b}_2 + (-\mathbf{b}_4) = (-\mathbf{b}_3)$. This non-uniqueness brings about an interesting feature of the dislocation network topology in b.c.c. crystals that is not present in the conventional dislocation networks consisting solely of binary junctions. In the latter, it is possible to trace every single line with $1/2\langle 111 \rangle$ Burgers vector through each $\langle 001 \rangle$ junction it enters. It is even possible to uniquely deconstruct the entire network into individual $1/2\langle 111 \rangle$ lines. However, the topology of dislocation networks containing 4-nodes is different in principle: it is now impossible to specify which of the four dislocations in a particular 4-node are the parents and which is the product, and to 'trace' a given $1/2\langle 111 \rangle$ line through the network. While it is still possible to deconstruct the network into constituent lines, there are a combinatorially large number of different ways to do this. The formation of multi-junctions results in the topological irreversibility or untraceability of dislocation networks.

Although the DD simulations provide credible support for the existence of multi-junctions, it is desirable to verify this finding with a discrete atomistic model that does not rely on the continuum theory of dislocations. Figure 2a shows the result of one such simulation in which three different dislocations (left-hand side of equation (1)) were introduced into a small fragment of the b.c.c. single crystal and then allowed to relax the lattice distortions produced by the inserted dislocations (see the Methods section for details). In the relaxed configuration (Fig. 2a), two distinct 4-nodes are instantly recognizable, as is the junction dislocation with the $1/2[111]$ Burgers vector (right-hand side of equation (1)). Taken together, the DD and the atomistic simulations appear to be convincing: multi-junctions should exist.

Yet, for definitive verification, we rely on transmission electron microscopy (TEM) of molybdenum single crystals. Figure 2b–d shows three different views of a single fragment of the dislocation network containing a binary junction and a 4-node. Four lines entering the 4-node are numbered from 1 to 4. A unique TEM signature of the symmetric 4-node is that, in certain contrast conditions, one of four dislocations must appear out of contrast while the other three lines remain visible. The appearance of a symmetric 4-node in each of the four TEM frames shown in Fig. 2c and d is unmistakable. Similar 4-node dislocation configurations were also found in other regions of the TEM foil, leading us to believe that their occurrence may not be rare.

As discussed above, multi-junctions appear to hold dislocations together more tightly than binary junctions. It is difficult to quantify this difference exactly because the stress required to unzip a given

junction depends on multiple factors, including dislocation line lengths and orientations, direction of applied stress and exactly how the junction is incorporated in the dislocation network. To assess qualitatively the relative holding power of binary and multi-junctions, we performed a large series of DD simulations in which both binary and multi-junctions are formed in the same geometries and subjected to the same straining conditions (see Supplementary Discussion 1 for the detailed results). Depending on the Burgers vectors and line orientations, the dislocations may mutually repel or attract each other. In cases when they attract, the lines either zip junctions or stay crossed¹². The crossed attractive configurations do not hold dislocations together appreciably and are destroyed by application of a small stress, whereas junction unzipping (Fig. 1c) necessitates a significantly higher stress.

The superior strength of multi-junctions compared to binary junctions manifests itself in several ways. First, the multi-junctions form and exhibit measurable strength over a wider range of initial line orientations than the binary junctions. Second, in the collision geometries where both binary and multi-junctions zip, the latter require much higher stress to unzip and release the dislocations. Finally, whereas the binary junctions could eventually be unzipped under all tested line and stress orientations, the multi-junctions were found to be indestructible across a wide range of line orientations and stress states. In such cases, rather than unzip and release the lines to

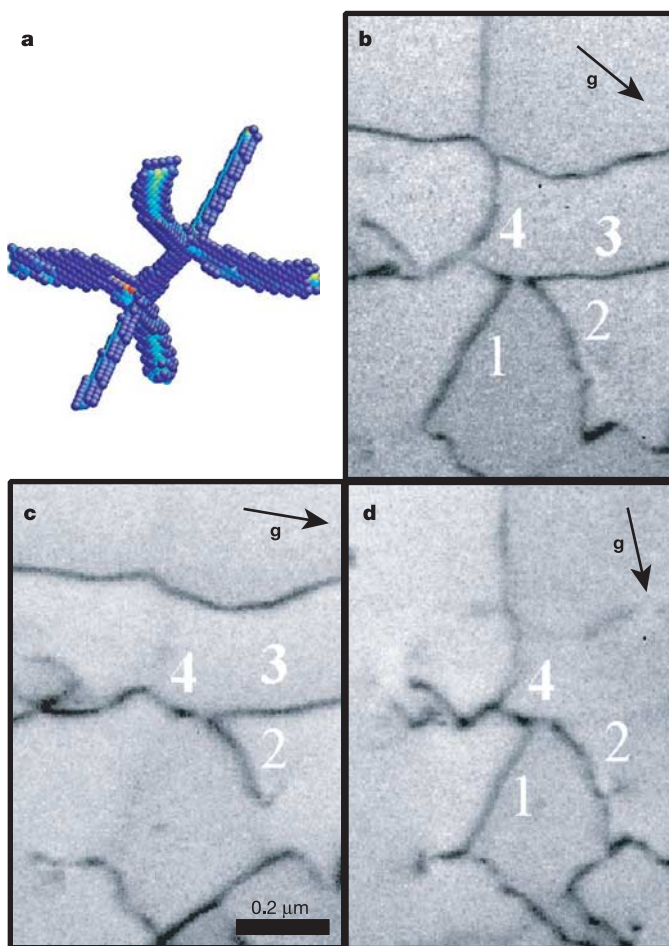


Figure 2 | Atomistic simulations and experiments confirm that multi-junctions exist in b.c.c. molybdenum. **a**, A multi-junction formed in an atomistic simulation. **b**, A TEM micrograph containing a symmetric 4-node. In this view all four dislocations (1–4) entering the multi-node are visible. **c**, View in which dislocation 1 becomes invisible. The length of the scale bar is $0.2\ \mu\text{m}$. **d**, Another view in which dislocation 3 is invisible.

glide under increasing stress (Supplementary Video 3), the multi-junction yields by repetitively emitting concentric dislocation loops and returning to its zipped configuration (Fig. 1f and Supplementary Video 4), thus forming a new regenerative dislocation source of the Frank–Read type¹⁵. In contrast, under no stress state was a binary junction observed to act as a regenerative dislocation source.

We now attempt to determine whether multi-junctions play a role in strain hardening, that is, when continued straining demands increasingly higher stress. The small-scale DD simulations reveal the great strength of multi-junctions and their propensity to act as regenerative sources and imply that these dislocation tangles may play an important part in strain hardening, but alone are inconclusive. At the same time, transmission electron micrographs have demonstrated the existence of multi-junctions in b.c.c. molybdenum and support a connection between multi-junctions and strain hardening, but because they were taken several days after the straining experiments, are not conclusive. Large-scale DD simulations present the opportunity to observe the formation of multi-junctions during straining and to quantify their effects on strain hardening, via *in situ* computational experiments.

The extreme computational cost of DD simulations has made it unfeasible to simulate dislocation ensembles large enough to compute strain hardening directly from the underlying motion and interactions of many dislocation lines. Recently, we developed a new DD code *ParaDiS* (for Parallel Dislocation Simulator) specifically designed to take full advantage of massively parallel supercomputers¹⁶. *ParaDiS* runs efficiently on two of the world's most powerful computers, Thunder and Blue Gene/L, both at the University of California Lawrence Livermore National Laboratory (http://www.llnl.gov/computing/hpc/resources/OCF_resources.html). Here *ParaDiS* enables direct simulations of large strains and strain hardening in statistically representative dislocation ensembles.

Figure 3 shows the data obtained in a series of three DD simulations of single crystal molybdenum subjected to uniaxial compression along two different directions (see also Supplementary Fig. 1). The simulation parameters were chosen to represent molybdenum single crystals at an elevated temperature of 450 K. In accordance with experiments⁸ (see also Supplementary Fig. 2), the

simulation of uniaxial straining along [001] direction exhibits a high rate of strain hardening, as shown by a pronounced slope of the stress–strain curve beyond the initial yield, whereas there is virtually no hardening exhibited in the [011] case (Fig. 3a). Because dislocations of at least three distinct Burgers vectors must be present to form multi-junctions, they are expected to form more frequently in the [001] straining simulation where all four distinct Burgers vectors of $1/2\langle 111 \rangle$ type are active and not in the [011] case where only two of the four are active. Observed both in simulations and in experiment, the strong contrast in the orientation dependence of strain hardening makes it tempting to assert that the difference is somehow related to multi-junctions. This assertion is further supported by the diminishing multiplication rate observed in the [011] case compared to the rapid and steady dislocation multiplication observed in the [001] case (Fig. 3b), and explained by the propensity of multi-junctions to form additional dislocation sources.

To test further the assertion that multi-junctions strongly influence the orientation dependence of strain hardening, we now perform a computational experiment that is impossible to reproduce in the laboratory. Specifically, we repeat the same [001] straining simulation but start with a modified initial configuration in which the Burgers vectors $\mathbf{b}_3$ and $\mathbf{b}_4$ are converted into $\mathbf{b}_1$ and $\mathbf{b}_2$, respectively, such that the total density of dislocations remains initially unchanged. By ‘doctoring’ the initial structure in such a way, we preclude any possibility of multi-junction formation during the course of this simulation. The resulting stress–strain and stress–density behaviours differ markedly from the full [001] straining simulation (with all four Burgers vectors included). Instead, the behaviour is similar to that observed in the [011] straining simulation (Fig. 3a and b). This observation further reinforces our assertion that the high hardening rate in straining along the high symmetry (for example, [001]) directions is related to the formation of the multi-junctions (see Supplementary Discussion 2).

To clarify how the multi-junctions affect strain hardening, we investigate the relationship between the evolving dislocation microstructure and the instantaneous flow stress. As shown in Fig. 3c, over the range of dislocation densities common to all three simulations, the flow stress appears to be determined by the total dislocation

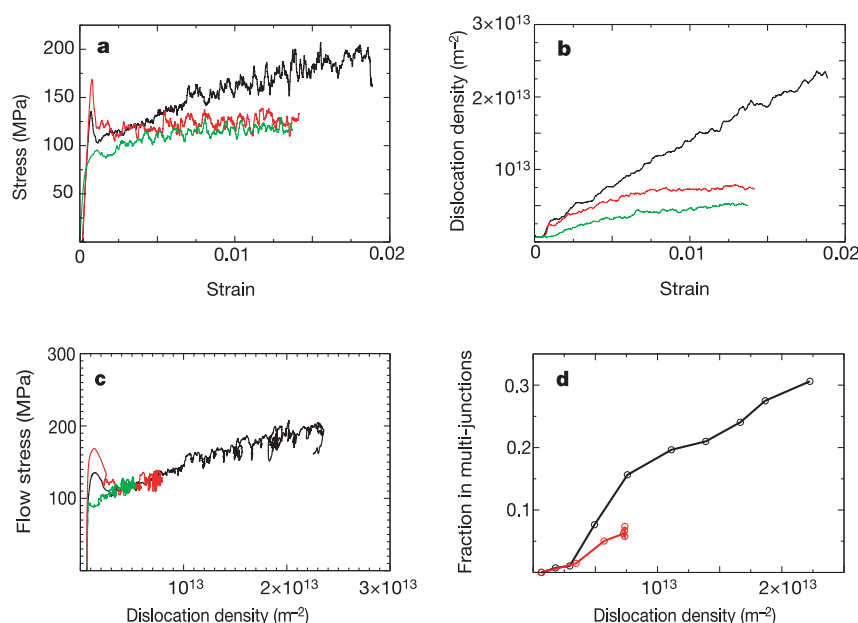


Figure 3 | The results of virtual straining experiments on b.c.c. molybdenum. The black lines correspond to the full [001] straining simulation, the red lines are for the [011] straining, and the green lines are for the ‘doctored’ [001] straining simulation in which two of the four

Burgers vectors are absent. **a**, Flow stress as a function of strain.

b, Dislocation line density as a function of strain. **c**, Flow stress versus total dislocation density. **d**, The fraction of lines involved in multi-junction configurations as a function of the total line density.

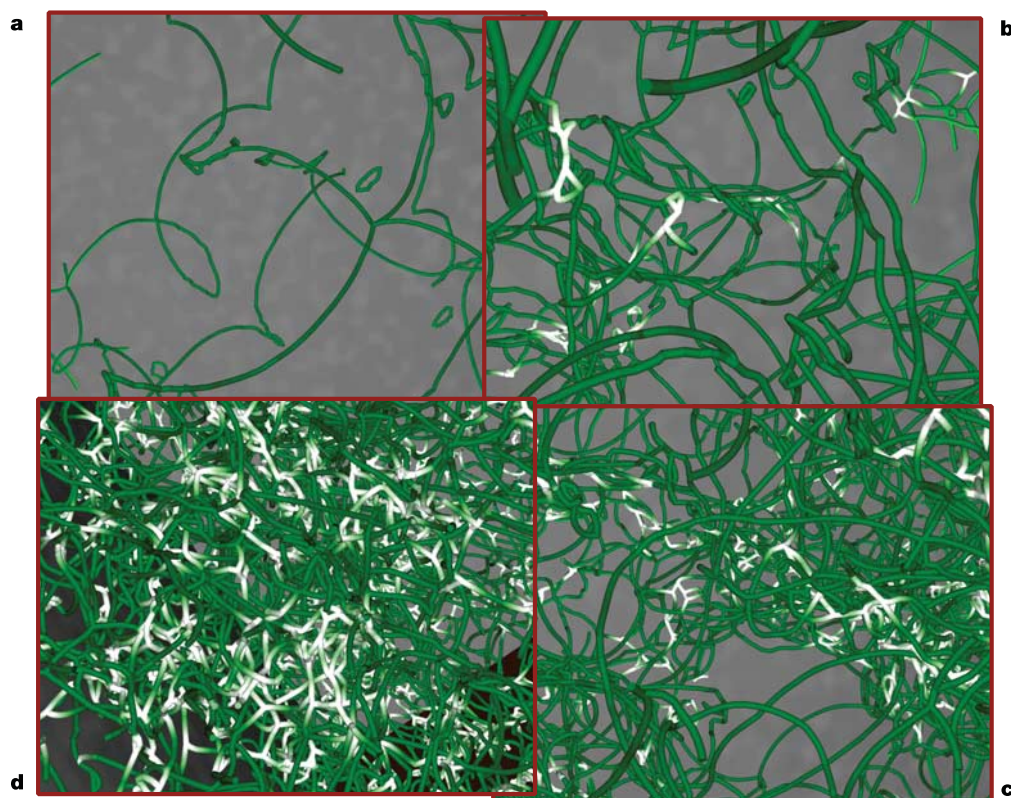


Figure 4 | Snapshots of dislocation network evolution obtained from a DD simulation of [001] straining. **a**, Initially, dislocation motion results in binary collisions only so that the network remains all green (see Methods section for a more detailed description of the colour scheme). **b**, Near the

yield strain ($\sim 0.2\%$), dislocations multiply and their collisions produce the first few multi-junctions (white). **c**, **d**, Continued dislocation multiplication results in increasingly frequent dislocation collisions leading to strain hardening and growth of the (white) sub-network of multi-junctions.

density alone, irrespective of the orientation of the tensile axis and the number of active Burgers vectors, suggesting that the weaker but more numerous binary interactions predominantly define the overall plastic strength. At the same time, the frequency of binary collisions is determined by the dislocation density that increases much more rapidly in the high-symmetry, [001] straining due to proliferation of new dislocation sources, multi-junctions.

As shown in Fig. 3d, the topological composition of the dislocation microstructure differs significantly for different straining directions at the same level of total dislocation density (see also Supplementary Fig. 3). Whereas in the [011] straining simulation both the total dislocation density and the fraction of lines involved in the multi-junctions saturate, the [001] straining simulation with the higher fraction of multi-junction configurations continues to evolve to considerably higher total densities and higher fractions of lines involved in the multi-junctions. Taken together, these observations indicate that the number of active Burgers vectors affects the rate of formation of new dislocation sources (multi-junctions), leading to significant differences in the accumulation of dislocation density and, thus, in the strain hardening rates.

In situ visual observations reveal that, even in the [001] straining simulations, the binary junctions form much more frequently than the multi-junctions. In fact, for as long as dislocations move on different planes they must intersect, making formation of binary junctions unavoidable. In contrast, multi-junctions form infrequently, mostly by attachment of a third line to an existing binary junction. At the same time, while the binary junctions are observed to dynamically unzip and reform elsewhere, the multi-junctions, once formed, are observed to endure. Zipping of new multi-junctions takes place preferentially near the existing ones (see Supplementary Video 5), gradually building up a sub-network of multi-junctions, a strong and mostly static backbone of the growing microstructure (Fig. 4).

It remains to be seen what role multi-junctions play in the intricate dislocation patterns formed during straining of high-symmetry crystals¹⁷: it seems that multi-junctions can serve as strong anchors for dislocation tangles, braids, walls, etc. As to whether multi-junctions form in crystals other than b.c.c., we predict that a variety of strong multi-junctions should exist in the face-centred-cubic and related high-symmetry crystals. Finally, theoretical analysis, DD simulations and TEM observations all suggest that dislocation tangles even more complex than ternary junctions exist, but are rare and their stability is likely to be marginal.

METHODS

Dislocation dynamics simulations. In a DD simulation, dislocations are represented by piece-wise straight segments interacting with each other according to the equations of the continuum elasticity theory^{18,19}. In a simulation, each dislocation segment moves with velocity $\mathbf{v} = M \cdot \mathbf{f}$ proportional to the net force $\mathbf{f}$ exerted on the segment by external loads and all other dislocation segments (here M is the mobility tensor). A single simulation step includes: (1) calculating the forces acting on the segments, (2) advancing the segments to new positions according to their velocities, and (3) performing changes in the line topology (connectivity) when collisions or node instabilities are encountered. The force on a dislocation segment is calculated as the negative derivative of the system's energy with respect to the segment's position. The elastic constants and dislocation mobility function were chosen to capture the behaviour of b.c.c. molybdenum above its athermal threshold (~ 450 K). The mobility was independent of the local line direction in the glide plane: that is, the plane containing both line direction and Burgers vector. Mobility in the direction normal to the glide plane was a small fraction (10^{-6}) of the glide mobility. Screw dislocations, whose line directions are parallel to their Burgers vectors, were free to glide in any plane containing their line direction. All small-scale DD simulations were conducted as though the configurations were in an infinite medium, whereas the larger-scale straining simulations were performed in a periodic cube $5 \mu\text{m}$ on the side.

Analysis of dislocation network topology. The entire network is comprised of

nodes and links. A node is where three or more lines merge together, and a link is any line segment connecting two nodes of the network. For the present analysis, we label any 3-node that bounds a regular binary junction as a 'normal' or N-node. Likewise, any 4-node formed by two dislocations crossing each other is also labelled an N-node. All other nodes are regarded as multi-nodes or M-nodes, including the symmetric 4-nodes shown in Fig. 1e as well as 3-nodes produced by dissociation of the symmetric 4-nodes. Three types of links can now be defined with respect to the types of nodes they connect: NN links, NM links and MM links. Here, notation NM is used for any link that connects an N-node to an M-node. To compute the fraction of lines involved in the multi-junctions shown in Fig. 3d, we summed the lengths of all MM links with half the lengths of all NM links and divided this sum by the total length of all links. The colour scheme used in Fig. 4 is as follows: all MM links are shown in white, whereas the colour of NM links is graded from white at the M-nodes to green at the N-nodes. All NN links, including binary junctions, are shown in green.

Atomistic simulation. The simulation volume was a small cube-shaped block of a perfect b.c.c. single crystal, 17 nm on each side. The initial geometry contained three dislocations with Burgers vectors $1/2[\bar{1}11]$, $1/2[1\bar{1}1]$, and $1/2[11\bar{1}]$ intersecting at the block centre. The atom positions inside the block were then relaxed to mechanical equilibrium using the conjugate gradient method and an interatomic interaction function for molybdenum²⁰. The atoms on the block surfaces were fixed throughout the simulation. To visualize crystal defects, only the atoms inside the block with energies exceeding the ideal bulk value by 0.095 eV are shown.

Experiment. The experiments involved three steps: (1) compression of a single-crystal molybdenum specimen to 1% total strain along the $[001]$ axis, (2) cutting and thinning the deformed specimen along the $(\bar{1}01)$ plane to obtain electron transparent foils, and (3) TEM observations using a set of reflection vectors $\mathbf{g}$ that can reveal multi-junctions. In the view shown in Fig. 2b the zone axis $\approx [\bar{1}01]$ and the diffraction vector $\mathbf{g} = [020]$, making all four dislocations entering the 4-node visible. The views in Fig. 2c and d were obtained using $\mathbf{g} = [\bar{1}21]$ and $\mathbf{g} = [121]$ which made lines $\mathbf{b}_1 = 1/2[111]$ and $\mathbf{b}_3 = 1/2[\bar{1}\bar{1}1]$ invisible owing to the $\mathbf{g} \cdot \mathbf{b} = 0$ condition. To access additional diffraction vectors, the specimen was tilted to a new zone axis ($\approx [\bar{2}01]$) making it possible to identify the Burgers vectors of two remaining dislocations in a similar manner: $\mathbf{b}_2 = 1/2[11\bar{1}]$ and $\mathbf{b}_4 = 1/2[\bar{1}11]$, respectively.

Received 14 November 2005; accepted 14 February 2006.

- Orowan, E. Zur kristallplastizität. *Z. Phys.* **89**, 605–659 (1934).
- Taylor, G. The mechanism of plastic deformation in crystal. Part I. Theoretical. *Proc. R. Soc. A* **145**, 363–404 (1934).
- Polanyi, M. Über eine Art Gitterstörung, die einen kristall plastisch machen konnte. *Z. Phys.* **89**, 660–664 (1934).
- Hirsch, P. B., Horne, R. W. & Whelan, M. J. Direct observations of the arrangement and motion of dislocations in aluminium. *Phil. Mag.* **1**, 677–684 (1956).
- Hirth, J. P. & Lothe, J. *Theory of Dislocations* 2nd edn, 9–13 (Wiley, New York, 1982).
- Saada, G. Sur le durcissement dû à recombinaison des dislocations. *Acta Metall.* **8**, 841–852 (1960).
- Franciosi, P., Berveiller, M. & Zaoui, A. Latent hardening in copper and aluminium single crystals. *Acta Metall.* **28**, 273–283 (1980).
- Guiu, F. & Pratt, P. L. The effect of orientation on the yielding and flow of molybdenum single crystals. *Phys. Status Solidi* **15**, 539–552 (1966).
- Hirth, J. P. & Lothe, J. *Theory of Dislocations* 2nd edn, 269–270 (Wiley, New York, 1982).
- Basinski, S. J. & Basinski, Z. S. in *Dislocations in Solids* Vol. 4 (ed. Nabarro, F. R. N.) 261–362, (North Holland, Amsterdam, 1979).
- Nabarro, F. R. N., Basinski, Z. S. & Holt, D. B. The plasticity of pure single crystals. *Adv. Phys.* **50**, 193–233 (1964).
- Madec, S. R., Devincre, B. & Kubin, L. P. From dislocation junctions to forest hardening. *Phys. Rev. Lett.* **89**, 255508 (2002).
- Hähner, P. & Zaiser, M. Dislocation dynamics and work hardening of fractal dislocation cell structures. *Mater. Sci. Eng. A* **272**, 443–454 (1999).
- Abraham, F. F. et al. Simulating materials failure by using up to one billion atoms and the world's fastest computer: Work hardening. *Proc. Natl. Acad. Sci. USA* **99**, 5783–5787 (2002).
- Frank, F. C. & Read, W. T. Multiplication processes for slow moving dislocations. *Phys. Rev.* **79**, 722–723 (1950).
- Bulatov, V. V. et al. Scalable line dynamics in ParaDiS. *Supercomputing* (<http://www.sc-conference.org/sc2004/schedule/pdfs/pap206.pdf>) (2004).
- Mughrabi, H. Dislocation wall and cell structures and long-range internal stresses in deformed metal crystals. *Acta Metall.* **31**, 1367–1379 (1983).
- Kubin, L. P. & Canova, G. R. The modelling of dislocation patterns. *Scr. Metall. Mater.* **27**, 957–962 (1992).
- Ghoniem, N. M., Huang, J. M. & Wang, Z. Q. Affine covariant-contravariant vector forms for the elastic field of parametric dislocations in isotropic crystals. *Phil. Mag. Lett.* **82**, 55–63 (2002).
- Finnis, M. W. & Sinclair, J. S. A simple empirical N-body potential for transition metals. *Phil. Mag. A* **50**, 45–55 (1984).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the US DOE Office of Basic Energy Sciences and the NNSA ASC program. We thank E. Chandler and C. Mailhot for their encouragement and unwavering support of the ParaDiS code development effort, D. Lassila for providing the purified molybdenum crystals, M. LeBlanc for performing the mechanical tests, R. Cook for graphic design and F. Abraham and G. Campbell for critical reading and editorial suggestions for the manuscript. This work was performed under the auspices of the US DOE by the Lawrence Livermore National Laboratory.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to V.V.B. (bulatov1@llnl.gov).

The twentieth century was the wettest period in northern Pakistan over the past millennium

Kerstin S. Treydte¹, Gerhard H. Schleser², Gerhard Helle², David C. Frank¹, Matthias Winiger³, Gerald H. Haug⁴ & Jan Esper¹

Twentieth-century warming could lead to increases in the moisture-holding capacity of the atmosphere, altering the hydrological cycle and the characteristics of precipitation¹. Such changes in the global rate and distribution of precipitation may have a greater direct effect on human well-being and ecosystem dynamics than changes in temperature itself^{2,3}. Despite the co-variability of both of these climate variables³, attention in long-term climate reconstruction has mainly concentrated on temperature changes^{4–8}. Here we present an annually resolved oxygen isotope record from tree-rings, providing a millennial-scale reconstruction of precipitation variability in the high mountains of northern Pakistan. The climatic signal originates mainly from winter precipitation, and is robust over ecologically different sites. Centennial-scale variations reveal dry conditions at the beginning of the past millennium and through the eighteenth and early nineteenth centuries, with precipitation increasing during the late nineteenth and the twentieth centuries to yield the wettest conditions of the past 1,000 years. Comparison with other long-term precipitation reconstructions indicates a large-scale intensification of the hydrological cycle coincident with the onset of industrialization and global warming, and the unprecedented amplitude argues for a human role.

The high mountain systems of Central Asia control atmospheric circulation patterns with stationary low and high pressure systems resulting from heating and cooling interactions with the lower troposphere, respectively⁹. Although there is increasing knowledge about regional temperature trends over the last millennium—showing warmth in Medieval times, cooling during the ‘Little Ice Age’ and recent warming^{10,11}—few investigations, restricted primarily to recent centuries and focusing mainly on monsoonal regions^{12–17}, address variations in precipitation. However, improving knowledge regarding long-term precipitation changes and potential large-scale

trends is of great importance for future predictions of regional and global hydrological cycles².

Our study sites include the western part of High Asia with the Karakorum and Himalayan mountains in northern Pakistan, surrounding the upper reaches of the Indus Valley and supplying the world’s largest irrigation network. This high mountain system interacts predominantly with westerly synoptic fronts and catalyses the formation of meridional troughs in the jet stream (Rossby waves). The climate is of ‘Mediterranean character’, with precipitation being highly variable during summer and at a maximum during winter and spring, resulting in extensive snow cover.

We have evaluated precipitation changes through four, annually resolved oxygen isotope ratio ($\delta^{18}\text{O}$) chronologies from juniper tree-ring cellulose (*Juniperus excelsa*, *J. turkestanica*). Three of these records cover the twentieth century and one extends back to AD 828. The latter was used to reconstruct precipitation variability over the past millennium.

Our study sites are located in three valleys south and north of the main Karakorum range (Fig. 1). They transect a southwest–northeast precipitation gradient with increasing rain-shadow effects northward through the mountain ranges. Three sites are situated near the upper timberline (>3,700 m above sea level, a.s.l.) and one near the lower timberline (2,900 m). According to local site ecology and vegetation cover classification, two sites were classified as cold/moist (‘Ram’ and ‘Bag-high’), one as warm/dry (‘Bag-low’), and one, containing the >1,000-year-old trees, as cold/dry (‘Mor’). Except for the Mor site, $\delta^{18}\text{O}$ measurements are confined to the period 1900–1998 (Supplementary Table 1). All sites inter-correlate significantly ($P < 0.001$) (Fig. 2a), suggesting common climatic forcing independent of potentially differing moisture sources, elevation and local ecological conditions. Similarities in the annual series are not limited to the

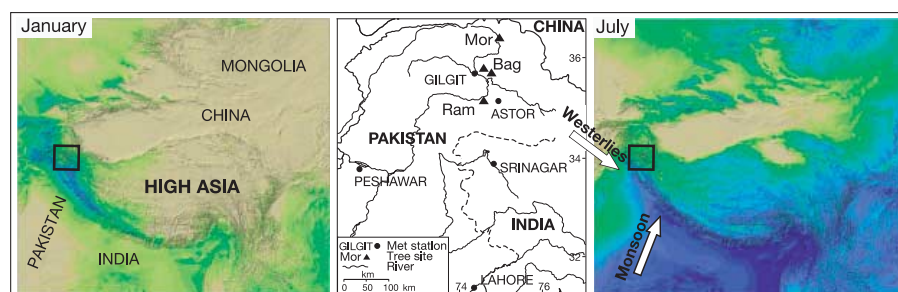


Figure 1 | Site locations and Central Asian average precipitation distribution⁹ in January and July. The study area is dominated by westerly synoptic fronts throughout the year (dark colours indicate high precipitation levels). ‘Bag’ includes one high (3,900 m a.s.l.) and one low

(2,900 m a.s.l.) elevation site. ‘Mor’ (3,900 m a.s.l.) contains >1,000-year-old living trees. The maps are reprinted from ref. 9 (<http://www.tandf.no/boreas>) with permission from Taylor & Francis.

¹Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland. ²Forschungszentrum Jülich GmbH, ICG-V, Leo-Brandt-Straße, D-52425 Jülich, Germany. ³University of Bonn, Meckenheimer Allee 166, D-53115 Bonn, Germany. ⁴GeoForschungsZentrum Potsdam, Telegrafenberg, C324, D-14473 Potsdam, Germany.

inter-annual (high-frequency) but also exist on decadal timescales. The range in 1900–1998 $\delta^{18}\text{O}$ mean values (28.6–31.1‰ between sites) is, however, related to absolute levels of local soil moisture availability, and hence to variable transpiration rates and leaf water enrichment¹⁸.

Correlation analyses using all $\delta^{18}\text{O}$ site records, their mean and variance-adjusted average (which we term Kara, derived from Karakorum), and regionalized climate data (REG; Supplementary Fig. 1) show no strong link to temperature variations but do correlate significantly with precipitation totals (Fig. 2b and Supplementary Fig. 3). This pattern is common between the sites, and has greatest similarities to mean annual precipitation (measured from the end of the vegetation period from the previous year to the end of the subsequent one: October–September), which is dominated by the winter half-year precipitation from October to March. Hence, tree-ring $\delta^{18}\text{O}$ serves as a proxy for the amount of precipitation, which at these elevations primarily falls as snow, providing the main source of water for the trees. This holds especially in spring, when early wood cells, which account for about 90% of the total ring width (roughly 80% of ring mass), are formed. The effect persists in part during summer, as much of the summer precipitation at high elevations falls as snow (but contributes less than 20% to the annual total). Snowmelt provides a more stable water source than rainfall with rapid surface runoff, which is supported through slightly positive correlations with summer temperatures. Potential effects of varying ratios of snow and rainfall could also affect the detected climate signal, but were not supported by calibration tests against instrumental data. The detected signal stems largely from the markedly depleted snowfall that dominates the annual precipitation regime^{19,20}. The negative relationship between precipitation totals and tree-ring $\delta^{18}\text{O}$ is strengthened through complex effects of varying humidity on plant physiology¹⁸. Simplified, moist

atmospheric conditions will reduce transpiration and therefore reduce ^{18}O water enrichment in needles. When incorporated into organic matter, this results in lower tree-ring $\delta^{18}\text{O}$ values¹⁸.

The system of snowfall, melt water supply, soil moisture and transpiration leaves a fingerprint on the $\delta^{18}\text{O}$ signal in cellulose, leading to an integrated precipitation signal over all seasons. $\delta^{18}\text{O}$ measurements of Kara, integrating all twentieth-century site chronologies, correlate with the October–September precipitation series (REG) at $r = 0.58$ (1898–1990, $P < 0.001$). In particular, the decadal-scale variations of increased precipitation (1930s, 1960s) and reduced precipitation are captured quite closely (Fig. 3a). Calibration and verification tests using the 1898–1944 ($r = 0.49$, $P < 0.001$) and 1944–1990 ($r = 0.69$, $P < 0.001$) periods suggest temporal stability of this relationship, despite the reduced replication, spatial representativity and lower quality of early instrumental data. Correlations using only Mor data are highly significant but slightly reduced, with r values of 0.42 and 0.62 over the 1898–1944 and 1944–1990 periods, respectively, indicating that increased signal quality is obtained with greater numbers of sites. Verification tests using reduction of error (RE) and coefficient of efficiency (CE) statistics²¹ yielded 0.36 (RE) and 0.24 (CE) for the recent (1944–1990), and 0.14 (RE) and –0.03 (CE) for the early (1898–1944) verification periods, indicating high reconstruction quality.

We addressed reconstruction complications due to possible age effects in the record by indicating deviations from the reconstruction using all data (shown in black in Fig. 3b) compared to reconstruction using only tree-ring data from biologically old rings during the 1264–1599 and 1010–1179 periods (red in Fig. 3b). We find that potential age-related biases, as reported for reconstructions based on ring width⁵, only slightly affect long-term climate reconstructions using $\delta^{18}\text{O}$ measurements. The peak values reconstructed for the 1350s and 1470s using only juvenile tree-rings are higher relative

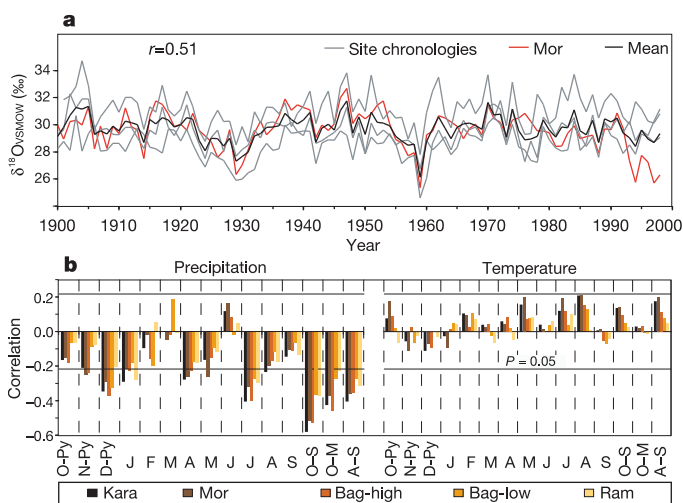


Figure 2 | $\delta^{18}\text{O}$ site chronologies and climate correlation. **a**, Site chronologies, including the most recent part of the millennial-long Mor record, together with the grand mean of all sites. r is the mean inter-series correlation. $\delta^{18}\text{O} = [((^{18}\text{O}/^{16}\text{O})_{\text{sample}} / (^{18}\text{O}/^{16}\text{O})_{\text{standard}}) - 1] \times 1,000$. **b**, Correlations between regionalized precipitation and temperature records (REG), the $\delta^{18}\text{O}$ site records, and the adjusted average, 'Kara'. REG was calculated by averaging station anomalies for temperature, and normalized indices for precipitation. Kara is the arithmetic mean of the site records, after adjusting the twentieth-century mean $\delta^{18}\text{O}$ values and their variance to Mor. Climate correlations are calculated for individual months from October of the previous year ('O-Py') to September of the following year ('S'), their mean ('O-S'), winter half-year (October–March, 'O-M') and summer half-year (April–September, 'A-S'). Correlations cover the full period of instrumental station data with a minimum replication > 2 (AD 1898–1990) (Supplementary Figs 1, 3). 95% significance levels are adjusted for lag – 1 autocorrelation.

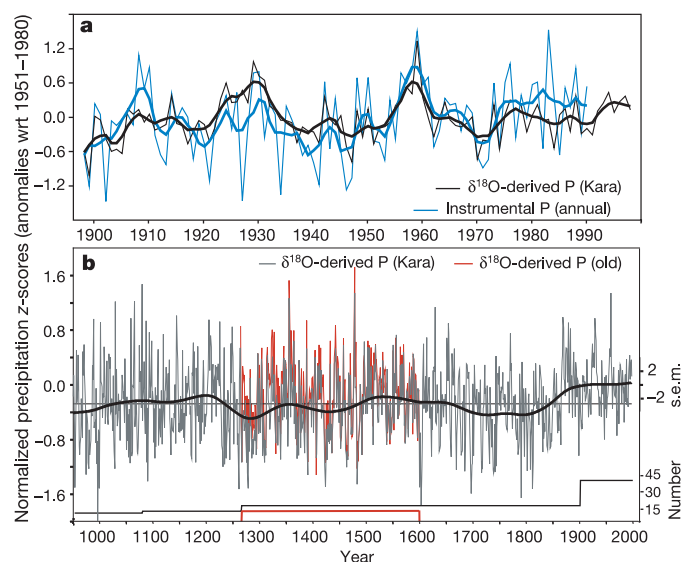


Figure 3 | $\delta^{18}\text{O}$ -derived precipitation reconstruction for northern Pakistan. **a**, Annual precipitation values (1898–1990) of instrumental and modelled data with respect to (wrt) the 1951–1980 mean. Smoothed curves are 5-year Kernel filters. **b**, Precipitation reconstruction using all tree-ring data (Kara) since AD 950. Proxy data were calibrated using a linear regression (1898–1990). Note that the raw $\delta^{18}\text{O}$ record is inverted here owing to the negative sign of the regression slope (Supplementary Fig. 4). The horizontal line is the overall mean of the reconstruction. For the period 1264–1599, maximum deviations are in red when using data from only old tree-rings (Supplementary Fig. 4 and Supplementary Notes). Replication of the reconstruction is three trees (6 cores) in AD 950. Long-term variations are highlighted using a 150-year spline. Uncertainty estimates for the low-frequency domain are indicated by two standard errors (see Methods).

to the continuous record integrating all data. Nonetheless, this age-related bias does not change the overall centennial-scale (low-frequency) pattern of reconstructed precipitation (Supplementary Fig. 4 and Supplementary Notes).

Our precipitation reconstruction over a millennium (Fig. 3b) reveals a cluster of peak precipitation periods during the late nineteenth and the twentieth century that are towards the upper limit of the distribution reconstructed over the past 1,000 years, with the majority of values before the twentieth century falling below the 1951–1980 mean. Wet periods of decadal length occur around 1200, 1350, 1500 and 1870; and dry periods occur before 1000 and around 1270, 1420, 1600 and 1720. The dry 1790s and 1890s are known as periods of severe famines in India¹². Tree-ring reconstructions from across High Asia confirm particularly intensified pluvial conditions in the twentieth century in India^{13,14}, Mongolia¹⁵ and Tibet¹⁶.

The spectral characteristics of the millennium-long $\delta^{18}\text{O}$ -derived precipitation reconstruction (Supplementary Fig. 5) show more low-frequency loading than is commonly reported for observational precipitation data²². This lower-frequency variation over the past

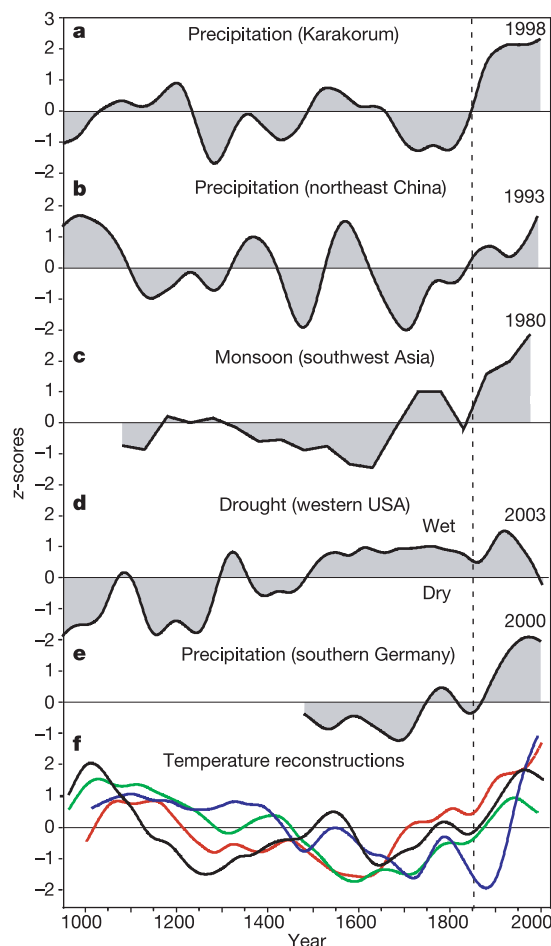


Figure 4 | Precipitation reconstruction for northern Pakistan and long-term precipitation variations for different regions in the northern hemisphere. **a**, Tree-ring $\delta^{18}\text{O}$ -derived reconstruction. **b**, Annual precipitation reconstruction (July–June) from tree-rings in northeast China¹⁷. **c**, Southwest Asian monsoon intensity from *Globigerina bulloides* in the Arabian sea¹². **d**, Drought reconstruction from tree rings in western USA²³. **e**, Spring–summer precipitation reconstruction from tree-rings in southern Germany²⁴. **f**, Regional to hemispheric temperature variations according to ref. 11 (red, western Central Asia), ref. 4 (blue), ref. 5 (black) and ref. 8 (green). Records are normalized over their individual periods and smoothed using 150-year splines. Numbers in **a–e** refer to the last year of the records, and the black dashed line to the shift from negative to positive precipitation anomalies in the Karakorum record.

1,000 years has trends similar to those found in long-term precipitation and drought reconstructions from synoptically different regions. Figure 4 shows the low-frequency components of our precipitation record (Fig. 4a) together with recently developed long-term precipitation reconstructions from semi-arid northeast China¹⁷, monsoonal southwest Asia¹², the western United States²³ and Germany²⁴. As these reconstructions represent rather different atmospheric circulation regimes, and considering that precipitation variations are spatially and temporally more heterogeneous than temperature variations, it is not surprising that these time series do not correlate on annual-to-decadal scales. However, for the lowest-frequency component, there is evidence for a large-scale trend towards pluvial conditions in the twentieth century, particularly in Eurasian reconstructions (Fig. 4). In the western United States, the recent trend in the second half of the twentieth century indicates a change towards drier conditions, related to Pacific sea surface temperature and El Niño/Southern Oscillation dynamics²³. Nevertheless, the overall long-term trend towards pluvial conditions in all records is consistent with a regional-to-large-scale temperature increase over the past 150 years^{4,5,8,11} (Fig. 4f), which is itself described to be at least partly forced by the increase in anthropogenic greenhouse gases during this time²⁵.

Globally observed annual precipitation trends indicate a twentieth-century increase of 9 mm over land areas, which is approximately 0.98% per decade²². There is evidence that warming leads to increases in the moisture-holding capacity of the atmosphere at a rate of about 7% K^{-1} , altering the hydrological cycle and characteristics of precipitation events, including their amount, frequency, intensity and duration. Depending on the feedback mechanisms included, similar or slightly larger increases are estimated for rainfall intensity¹. However, best estimates for total rainfall increases are smaller, falling within the range of ~ 1.0 – 3.4% K^{-1} (refs 2, 26). These differences suggest interplay between rainfall characteristics and the dependence of precipitation totals on total available energy, which is decreased by the increased ability of the troposphere to radiate away latent heat released by precipitation¹. A compilation of coupled ocean–atmosphere and atmosphere-only model runs shows that precipitation decreases at subtropical latitudes, and increases at high latitudes, around the Equator and across much of Asia, consistent with the evidence presented here. Overall, a 6% increase in global mean precipitation between the 1981–2000 and 2081–2100 periods is predicted from these models²⁷.

The controls on past and future precipitation patterns and levels are certainly not exclusively related to temperature, but reflect complex circulation and thermodynamic effects with great regional and temporal variations. Nevertheless, our study identifies large-scale similarities in precipitation variation on centennial timescales despite dissimilarity in the higher-frequency domain. We suggest that an unprecedented twentieth-century intensification of the hydrological cycle in western Central Asia has already occurred. It seems that this finding is not only restricted to summer conditions and monsoonal regions, but rather is of a broader seasonal and spatial extent. Moreover, the structure of our record suggests that this climatic change is not simply a temporary fluctuation, but rather a secular trend. If this change were brought about by anthropogenic forcing of climate, as seems most plausible, then further changes can be expected in the near future, and high- and lowland ecological and economic systems in Pakistan and India will be forced to adapt.

METHODS

Sample preparation. Our sampling strategy was based on considering only sites with steep slopes ($>35^\circ$) and trees growing in shallow and well-drained soils, to minimize potential effects of long-standing, depleted soil and ground water. The sampling design in high-elevation environments greatly amplifies the influence of melt water on the tree-ring $\delta^{18}\text{O}$ signal. This influence dominates over other potential signals, such as temperature-dependent fractionation effects that are not found in the Karakorum junipers.

At all sites 12 to 20 trees were sampled, with four cores taken per tree. Ring widths were measured using a semi-automated RinnTech system with a

resolution of 0.01 mm, crossdated using standard procedures, and 5–7 trees per site (two cores per tree) were chosen for isotope analysis. Criteria for sample selection were low numbers of missing rings and regular ring boundaries. To facilitate analysis, we pooled most of the tree-ring material before cellulose extraction. At the Mor site, 14 cores from seven trees were pooled to reach back to AD 828. At the other sites, we pooled tree-rings from 4–5 trees²⁸. Single-tree measurements were also undertaken during selected periods. Signal strength analysis using the ‘expressed population signal’ (EPS)²⁹ indicates that a minimum of 3.5 series would be satisfactory to develop a (population-) representative site record, capturing 85% of the variance of the theoretical, infinite population chronology.

Tree-rings were separated with a scalpel and samples were ground using an ultracentrifugal mill (Retsch ZM1). Cellulose was extracted and cellulose samples (250 µg) were pyrolyzed to CO at 1,080 °C using a Carlo Erba 1500 elemental analyser (CE Instruments) interfaced with an Optima IRMS (Micromass Ltd). $\delta^{18}\text{O}$ values are referenced to Vienna standard mean ocean water (VSMOW). Overall analytical precision, estimated from periodic standard deviation calculations using a commercial cellulose standard (FLUKA), is $\pm 0.3\text{‰}$.

Meteorological data setup. We screened monthly and seasonal temperature and precipitation data for 30 meteorological stations. Five stations were selected for calibration purposes based on the distance to the tree sites, number of missing values, and homogeneity and period of measurements. Two of these stations, Gilgit (GIL, 1,460 m a.s.l.) and Astor (AST, 2,166 m a.s.l.), are located in the study region near the tree sites (Fig. 1), but their records are rather short (GIL = 48 years, AST = 36 years). Srinagar (SRI, 1,587 m a.s.l.), is the closest station with data spanning more than 100 years (AD 1898–1998), and correlates significantly with GIL and AST. The Peshawar (PES, 360 m a.s.l., AD 1864–1990) and Lahore stations (LAH, 214 m a.s.l., 1864–1990), located at a greater distance to the tree sites but have data for >100 years, correlate weakly with the inner-mountainous stations GIL and AST but significantly with SRI. PES and LAH were therefore used for better replication and to reinforce regional climate conditions, particularly in the early period (Supplementary Table 2).

For the development of a regional mean record (REG), temperature means were calculated by averaging monthly anomalies with respect to the 1951–1980 mean. To minimize the effect of significantly differing means and variances, regional precipitation means were calculated by averaging standardized values; that is, monthly data were normalized with respect to the lengths of individual station measurements (Supplementary Fig. 1). These z-scores were not converted back to ‘absolute’ values, because of the dependence of precipitation totals on elevation, making the choice of an ‘absolute’ rainfall level rather arbitrary. The 1898–1990 interval, which is covered by at least two stations, was used for calibration.

Uncertainty estimation. For the long reconstruction, 95% confidence intervals for the annual scale were estimated using twice the r.m.s. error derived from a ‘leave-one-out’ cross-validation over the entire 1898–1990 instrumental period (2 s.e.m. = 1.17). Confidence limits based on the r.m.s.e. of the verification data from the split period calibration/verification trials were similar, although about 1% and 10% larger for the early and late verification period data, respectively, demonstrating the benefit of using more data for model calibration. Some additional uncertainty in the reconstruction model may be related to poor early instrumental records and uncertainty in the statistical model definition due to a necessarily limited calibration interval. The short instrumental period does not allow for calibration/verification of centennial-scale variability. As no direct statistical calibration of centennial-scale variability is possible, the regression model is assumed to be frequency-independent. However, we estimated the standard error for the smoothed data by multiplying the square root of the unexplained variance by the standard deviation of the target instrumental data (2 s.e.m. = 0.22). At these wavelengths, this calculation results in a more conservative estimate than the methods for timescale-dependent error bars described earlier³⁰. Note that these confidence limits do not consider declines in the proxy quality back in time or potential age-trends in the isotope data, and assume the relationship between the proxy and instrumental data to be temporally stable. The standard error in Fig. 3 is expressed relative to the 1898–1998 mean.

Received 22 August 2005; accepted 23 March 2006.

1. Trenberth, K. E., Dai, A., Rasmussen, R. M. & Parsons, D. B. The changing character of precipitation. *Bull. Am. Meteorol. Soc.* **84**, 1205–1217 (2003).
2. Intergovernmental Panel on Climate Change, *Climate Change 2001: The Scientific Basis* (IPCC, Geneva, 2001).
3. Trenberth, K. E. & Shea, D. J. Relationships between precipitation and surface temperature. *Geophys. Res. Lett.* **32**, 10.1029/2005GL022760 (2005).
4. Mann, M. E., Bradley, R. S. & Hughes, M. K. Northern Hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophys. Res. Lett.* **26**, 759–762 (1999).
5. Esper, J., Cook, E. R. & Schweingruber, F. H. Low-frequency signals in long

chronologies for reconstructing past temperature variability. *Science* **295**, 2250–2253 (2002).

6. Briffa, K. R. & Osborn, T. J. Blowing hot and cold. *Science* **295**, 2227–2228 (2002).
7. Esper, J., Frank, D. C. & Wilson, R. J. S. Climate reconstructions: Low-frequency ambition and high-frequency ratification. *Eos* **85**, 113–119 (2004).
8. Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M. & Karlen, W. Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data. *Nature* **433**, 613–617 (2005).
9. Boehner, J. General climatic controls and topoclimatic variations in Central and High Asia. *Boreas* (in the press).
10. Bao, Y., Braeuning, A., Johnson, K. R. & Yafeng, S. General characteristics of temperature variation in China during the last two millennia. *Geophys. Res. Lett.* **29**, 10.1029/2001GL014485 (2002).
11. Esper, J., Frank, D. C., Wilson, R. J. S., Büntgen, U. & Treydte, K. Uniform growth behavior among central Asian low and high elevation juniper tree sites. *Mount. Res. Dev.* (in the press).
12. Anderson, D. M., Overpeck, J. T. & Gupta, A. K. Increase in the Asian southwest monsoon during the past four centuries. *Science* **297**, 596–599 (2002).
13. Hughes, M. K. in *Climate Since AD 1500* (eds Bradley, R. S. & Jones, P. D.) 415–431 (Routledge, London, 1995).
14. Singh, J. & Yadav, R. R. Spring precipitation variations over the western Himalaya, India, since AD 1731 as deduced from tree-rings. *J. Geophys. Res.* **110**, 10.1029/2004JD004855 (2005).
15. Pederson, N., Jacoby, G. C., D’Arrigo, R. D., Cook, E. R. & Buckley, B. Hydrometeorological reconstructions for Northeastern Mongolia derived from tree-rings: 1651–1995. *J. Clim.* **14**, 872–881 (2001).
16. Bräuning, A. & Mantwill, B. Summer temperature and summer monsoon history on the Tibetan Plateau during the last 400 years recorded by tree-rings. *Geophys. Res. Lett.* **32**, 10.1029/2004GL020793 (2005).
17. Sheppard, P. R. et al. Annual precipitation since 515 BC reconstructed from living and fossil juniper growth of northeastern Qinghai Province, China. *Clim. Dyn.* **23**, 869–881 (2004).
18. Roden, J. S., Lin, G. & Ehleringer, J. R. A mechanistic model for interpretation of hydrogen and oxygen isotope ratios in tree-ring cellulose. *Geochim. Cosmochim. Acta* **64**, 21–35 (2000).
19. Rozanski, K., Araguas-Araguas, L. & Gonfiantini, R. Relation between long-term trends of oxygen-18 isotope composition of precipitation and climate. *Science* **258**, 981–985 (1992).
20. Tang, K. & Feng, X. The effect of soil hydrology on the oxygen and hydrogen isotopic compositions of plants’ source water. *Earth Planet. Sci. Lett.* **185**, 355–367 (2001).
21. Cook, E. R., Briffa, K. R. & Jones, P. D. Spatial regression methods in dendroclimatology: a review and comparison of two techniques. *Int. J. Climatol.* **14**, 379 (1994).
22. New, M., Todd, M., Hulme, M. & Jones, P. Precipitation measurements and trends in the twentieth century. *Int. J. Climatol.* **21**, 1889–1922 (2001).
23. Cook, E. R., Woodhouse, C., Eakin, C. M., Meko, D. M. & Stahle, D. W. Long-term aridity changes in the western United States. *Science* **306**, 1015–1018 (2004).
24. Wilson, R. J. S., Luckman, B. H. & Esper, J. A 500-year dendroclimatic reconstruction of spring/summer precipitation from the lower Bavarian forest region, Germany. *Int. J. Climatol.* **25**, 611–630 (2005).
25. Tett, S. F. B. et al. The impact of natural and anthropogenic forcings on climate and hydrology since 1550. *Clim. Dyn.* (submitted).
26. Allen, M. R. & Ingram, W. J. Constraints on future changes in climate and the hydrological cycle. *Nature* **419**, 224–232 (2002).
27. Emori, S. & Brown, S. J. Dynamic and thermodynamic changes in mean and extreme precipitation under changed climate. *Geophys. Res. Lett.* **32**, 10.1029/2005GL023272 (2005).
28. Leavitt, S. W. & Long, A. Sampling strategy for stable isotope analysis of tree-rings in pine. *Nature* **311**, 145–147 (1984).
29. Wigley, T. M. L., Briffa, K. R. & Jones, P. D. On the average of correlated time series, with applications in dendroclimatology and hydrometeorology. *J. Clim. Appl. Meteorol.* **23**, 201–213 (1984).
30. Briffa, K. R. et al. Tree-ring width and density data around the Northern Hemisphere: Part 1, local and regional climate signals. *Holocene* **12**, 737–757 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to K. Briffa, E. Cook, E. Hendy, R. Wilson, M. Gagen, J. Waterhouse, M. Gumpert and U. Büntgen for comments and suggestions. We thank S. Andres, B. Kammer, W. Laumer, G. Reiss, M. Schrimpf and C. Welscher for laboratory assistance, and C. Welscher, M. Gumpert and A. Shafgat for logistical support in the field. This research was funded by the German Federal Ministry for Education and Research, the German Science Foundation Schl 3-1, the European Union (ALPIMP and ISONET) and the Swiss National Science Foundation (NCCR-Climate).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.S.T. (kerstin.treydte@wsl.ch).

Lamprey-like gills in a gnathostome-related Devonian jawless vertebrate

Philippe Janvier^{1,2}, Sylvain Desbiens³, Jason A. Willett³ & Marius Arsenault³

So far, the Palaeozoic fossil jawless vertebrates have not provided any direct evidence for the organization of the gills, apart from vague impressions—supposedly left by gill filaments—on the bony surface of the gill chamber in certain armoured forms or ‘ostracoderms’ (for example, osteostracans and heterostracans)¹. The latter are currently regarded as more closely related to the living jawed vertebrates (crown gnathostomes) than to the living jawless vertebrates (hagfish and lampreys, or cyclostomes)^{2,3}. Here we report the first direct evidence for the position of the gill filaments—possibly supported by gill rays—enclosed by gill pouches in a 370-million year (Myr)-old jawless vertebrate, *Endeiolepis*, from the Late Devonian fossil fish site of Miguasha, Quebec, Canada. This extinct jawless fish has much the same gill organization as living lampreys, although it possesses an unusually large number of gill pouches—a condition unlike that in any extant vertebrates and that raises questions about gill development⁴. *Endeiolepis* is currently regarded as a close relative of anaspids³, a group of 410–430-Myr-old ‘ostracoderms’. Assuming that current vertebrate phylogeny is correct, this discovery demonstrates that pouches enclosing the gills are primitive for vertebrates, but have been subsequently lost in jawed vertebrates.

The Late Devonian Escuminac Formation at Miguasha is renowned for the outstanding preservation of its fossil fishes, which sometimes show soft-tissue preservation⁵. This site notably yields two closely similar, soft-bodied fishes, *Euphanerops longaevus* (Fig. 1a) and *Endeiolepis aneri* (Fig. 1b–e), thought to be anaspids in which the dermal skeleton (the scales and superficial head bones) was supposedly reduced or lost⁶, and regarded as possible intermediates between anaspids and living lampreys^{1,7}.

These two ‘naked anaspids’ are generally preserved as tarry imprints and yield little anatomical information⁸. However, some specimens of *Euphanerops* have internal calcified structures thought to be mineralized cartilages⁹. *Euphanerops* also has a peculiar ladder-shaped imprint that extends ventrally from the anterior limit of the head to the anus, and is thought to be an unusually elongated branchial skeleton (Fig. 1a), similar to the ‘branchial basket’ of lampreys¹⁰, but comprising about 30 gill bars (instead of only seven in lampreys)^{7,8}.

New data about the structure of the branchial apparatus of these jawless fishes recently came from the second of these ‘naked anaspids’ from Miguasha, *Endeiolepis*, which is known from specimens that generally have less anatomical details than *Euphanerops*, but can readily be identified by a peculiar scalloped impression extending along its ventral margin and long thought to be a paired series of thin, poorly mineralized ‘ventrolateral scales’^{6,8} (Fig. 1b, c). The examination of a large number of *Endeiolepis* specimens showed that this impression, when three-dimensionally preserved, is in fact an elongated, sleeve-shaped structure (Fig. 1e), whose outline in lateral view exactly matches that of the branchial basket imprint

of *Euphanerops* (Fig. 1a). This suggests that the two structures are the same, preserved in different ways, and that *Euphanerops longaevus* and *Endeiolepis aneri* may well be the same fish (see the Supplementary Notes).

In some specimens of *Endeiolepis*, the surface of the internal cast of the branchial basket shows horizontal grooves, which suggest impressions of some gill-related structure (Fig. 1d). But a single specimen is fortuitously preserved in a way that allows a more detailed reconstruction of the branchial apparatus and gill filaments. In this large but incomplete specimen (Fig. 2), the branchial basket has been split open longitudinally during decay, thereby avoiding the overprinting of the left and right branchial structures on each other during fossilization, and has become embedded in a fine-grained and clay-like sediment, which provides an accurate natural cast of the branchial apparatus (Fig. 2a, b). It provides evidence for gill pouches that have been filled by sediment, and appear as a series of internal casts in the form of vertically elongated blocks separated by forwardly dipping oblique clefts (Fig. 2c and Supplementary Fig. 1a). The surface of these clefts has impressions of numerous parallel ridges and dark rods (Fig. 2d) that bear some resemblance to the gill rays and gill filament imprints of a jawed fish—the lobe-finned *Eusthenopteron*¹¹ (also from Miguasha)—and even more so to the preserved gills of a Late Devonian shark from the United States of America¹². This is the first direct evidence for the position and organization of the gill filaments in a fossil jawless vertebrate.

Although these impressions are clearly those of the gills, it remains unclear whether they are the actual impressions of the soft tissues of the gill filaments, of their afferent branchial arteries¹⁰, or of cartilaginous gill-ray-like structures. Contrary to the jawed vertebrates, living jawless fishes have no gill rays; however, possible gill-ray-like structures (though not necessarily enclosed by pouches) occur in the myllokunmingiids and yunnanozoans, two presumed stem vertebrates from the Lower Cambrian of Chengjiang, China^{13,14}, and this suggests that they may be a primitive character for vertebrates.

Gill pouches were hitherto known with certainty only in the living cyclostomes and their fossil representatives, but their presence has long been extrapolated, without any direct evidence, to all fossil jawless vertebrates^{15,16}. Although they were certainly present in *Endeiolepis*, it remains undecided whether this was also the case for other ‘ostracoderms’, such as osteostracans and heterostracans.

Anaspids, in which *Euphanerops* and *Endeiolepis* are classically included, were once considered as either ancestral, or closely related, to lampreys^{17,11}. This was mainly based on overall resemblance, notably their downwardly tilted caudal fin (Fig. 1a–c) and the presence of a dorsal median ‘nostril’ (nasohypophysial opening). In current vertebrate phylogenies (see Supplementary Fig. 2), anaspids fall among stem gnathostomes^{2,3}, and these resemblances are regarded as either shared primitive characters, or convergences.

¹UMR 5143, CNRS, Muséum National d'Histoire Naturelle, CP38, 57 rue Cuvier, 75231 Paris Cedex 05, France. ²The Natural History Museum, Cromwell Road, London SW7 5BD, UK. ³Parc National de Miguasha, CP183, Nouvelle, Quebec G0C 2E0, Canada.

Assuming that *Endeiolepis* is either a 'naked anaspid' or the sister group to anaspids, then gill pouches (long regarded as unique specialization of the cyclostomes), and possibly gill rays, would represent a primitive condition, lost in jawed vertebrates. The loss of gill pouches and coeval development of jaws have triggered important functional changes, notably the transition from passive to active inspiration, through the involvement of jaws in ventilation¹⁷.

In embryonic development, the gills are derived from the endoderm in hagfishes and lampreys, and the ectoderm in living gnathostomes. Which of these conditions is primitive for vertebrates remains debated, depending on whether cyclostomes are considered mono- or paraphyletic. The gill development in *Endeiolepis* will be unknown

forever. However, assuming that *Endeiolepis* is a stem gnathostome with gill pouches does not preclude the possibility that it had an intermediate condition, with partly septate gills that received some ectodermal contribution.

The very large number of gill pouches (or polybranchic condition) and their posterior extension to the anus in *Endeiolepis* and *Euphanerops* has no equivalent in living vertebrates and is likely to be a specialization. It may have to do with the regulation of oxygen supply in vertebrates with passive inspiration¹⁸ and living in a poorly oxygenated environment, as was the case in the Escuminac Formation

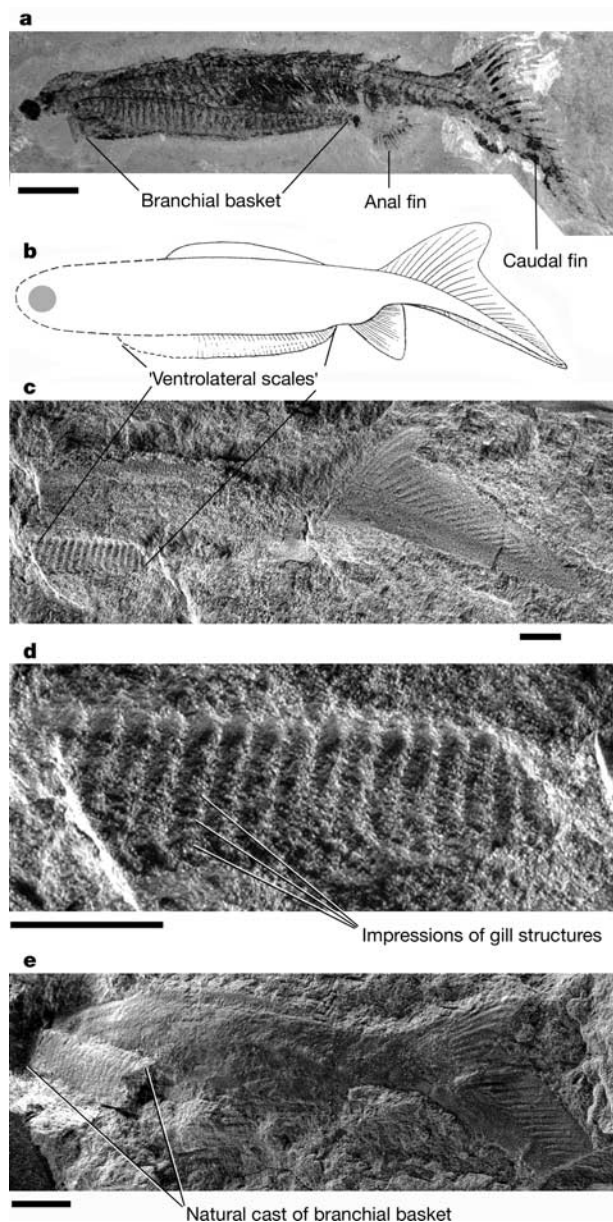


Figure 1 | The two 'naked anaspids' from the Upper Devonian of Miguasha, Quebec, Canada. **a**, *Euphanerops longaevus*, preserved as an imprint (Musée d'Histoire Naturelle, Miguasha, MHNM 01-02). **b–e**, *Endeiolepis aneri*. **b**, As classically reconstructed in lateral aspect, with paired 'ventrolateral scale' series (from refs 6, 8). **c**, **d**, Specimen showing the lateral impression of the branchial basket formerly thought to be 'ventrolateral scales' (MHNM 01-78) (**c**), and detail view showing impressions of gill structures (**d**). **e**, Specimen showing a three-dimensionally preserved natural cast of the branchial basket (MHNM 01-29A). Scale bar, 10 mm.

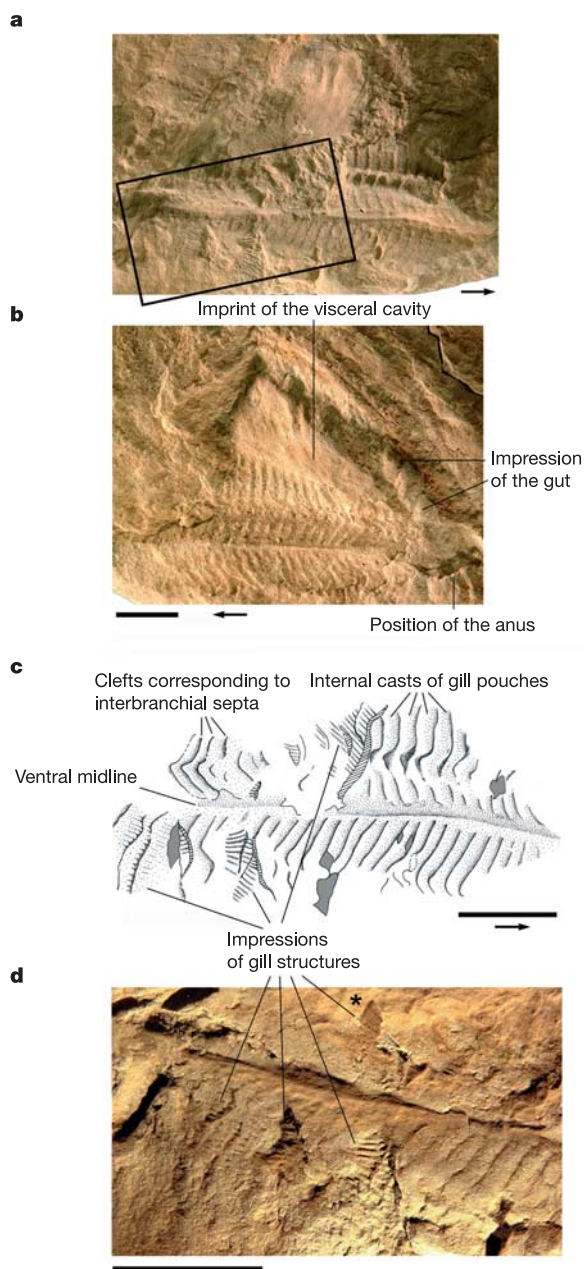


Figure 2 | *Endeiolepis aneri* (MHNM 01-154). **a**, **b**, Part and counterpart of a specimen showing the natural cast of the gill pouches of both sides of the branchial apparatus (split open during decay). **c**, Explanatory sketch of the part in **a** (damaged areas in grey; sections of broken gill-pouch casts obliquely hatched). **d**, Detail of the gill impressions (area framed in **a**); asterisk indicates where one gill-pouch cast has been removed to expose the surface of a cleft showing the corrugations and dark rods, which are imprints of either gill filaments, branchial arteries, or gill-ray-like structures. Arrows point forward. Scale bar, 10 mm.

in Late Devonian times¹⁹. Among living vertebrates, a moderately polybranchic condition is known only in certain hagfishes, which possess up to 15 gill pouches; however, the head shield of galeaspids, an extinct group of 430–370-Myr-old ‘ostracoderms’, has up to 45 pairs of depressions that are thought to have housed the gills. As there is no living proxy for such polybranchic vertebrates (apart from a superficial resemblance to the gill-less pharyngeal slits of the lancelet, *Branchiostoma*, which is not a vertebrate), it was assumed that these small and crowded gill chambers only contained vestigial gills, reduced to a mere folded respiratory epithelium⁴. A surprising result of this finding is the relatively large size of the gills in such a polybranchic vertebrate. Thus, *Endeiolepis* possessed a battery of about thirty pairs of normal-sized gills.

The development of the branchial skeleton in such cases of extreme polybranchic conditions may be a challenge for developmental biologists, whose experimental fish models are essentially lampreys and gnathostomes. Although the mechanisms involved in pharyngeal pouch development are barely elucidated^{20,21}, the condition in *Endeiolepis*, at any rate, suggests that the anterior endoderm could induce the formation of an almost unlimited number of pharyngeal pouches, as long as these could remain functional, probably at the cost of profound modifications in the position of other organs, such as the heart and digestive tract. Fossil gnathostome-related jawless vertebrates now demonstrate that the assembly of the gnathostome body plan was not a burst of anatomical innovations, and that lamprey-like gills could persist in fishes that already possess some gnathostome features, long before the rise of jaws²².

Received 29 August; accepted 21 November 2005.

- Janvier, P. *Early Vertebrates* (Oxford Univ. Press, Oxford, 1996).
- Donoghue, P. C. J., Forey, P. L. & Aldridge, R. J. Conodont affinity and chordate phylogeny. *Biol. Rev.* **75**, 191–251 (2000).
- Donoghue, P. C. J. & Smith, M. P. The anatomy of *Turinia pagei* (Powrie) and the phylogenetic status of the Thelodonti. *Trans. R. Soc. Edinb. (Earth Sci.)* **92**, 15–37 (2001).
- Janvier, P. in *Recent Advances in the Origin and Early Radiation of Vertebrates* (eds Arratia, G., Cloutier, R. & Wilson, M. V. H.) 29–52 (Dr Friedrich Pfeil, Munich, 2004).
- Arsenault, M., Desbiens, S., Janvier, P. & Kerr, J. in *Recent Advances in the Origin and Early Radiation of Vertebrates* (eds Arratia, G., Cloutier, R. & Wilson, M. V. H.) 439–454 (Dr Friedrich Pfeil, Munich, 2004).
- Stensiö, E. A. A new anaspids from the Upper Devonian of Scaumenac Bay in Canada, with remarks on the other anaspids. *K. Svenska Vetensk. Akad. Handl.* **18**, 1–25 (1939).
- Arsenault, M. & Janvier, P. in *Early Vertebrates and Related Problems of Evolutionary Biology* (eds Chang, M.-M., Liu, Y.-H. & Zhang, G.-R.) 19–37 (Science Press, Beijing, 1991).
- Janvier, P. in *Devonian Fishes and Plants of Miguasha, Québec, Canada* (eds Schultze, H. P. & Cloutier, R.) 134–140 (Dr Friedrich Pfeil, Munich, 1996).
- Janvier, P. & Arsenault, M. Calcification of early vertebrate cartilage. *Nature* **417**, 609 (2002).
- Marinelli, W. & Strenger, A. *Vergleichende Anatomie und Morphologie der Wirbeltiere. I. Lampetra fluviatilis* (L.) (Franz Deuticke, Vienna, 1954).
- Jarvik, E. *Basic Structure and Evolution of Vertebrates* Vol. 1 (Academic, London, 1980).
- Maisey, J. G. Visceral skeleton and musculature of a Late Devonian shark. *J. Vert. Paleont.* **9**, 174–190 (1989).
- Hou, H. G., Aldridge, R. J., Siveter, D. J., Siveter, D. J. & Feng, X.-H. New evidence on the anatomy and phylogeny of the earliest vertebrates. *Proc. R. Soc. Lond. B* **269**, 1865–1869 (2002).
- Mallatt, J. & Chen, J. Fossil sister-group of craniates: predicted and found. *J. Morphol.* **258**, 1–31 (2003).
- Watson, D. M. S. A consideration of ostracoderms. *Phil. Trans. R. Soc. Lond. B* **238**, 1–15 (1954).
- Stensiö, E. *Traité de Paléontologie* Vol. 4 (Masson, Paris, 1964).
- Mallatt, J. Ventilation and the origin of jawed vertebrates: a new mouth. *Zool. J. Linn. Soc.* **117**, 329–404 (1996).
- Massabuau, J. C. From low arterial- to low tissue-oxygenation strategy. An evolutionary theory. *Respir. Physiol.* **128**, 299–315 (2001).
- Prichonnet, G., Di Vergilio, M. & Chidiac, Y. in *Devonian Fishes and Plants of Miguasha, Québec, Canada* (eds Schultze, H. P. & Cloutier, R.) 23–36 (Dr Friedrich Pfeil, Munich, 1996).
- Graham, A. & Smith, A. Patterning the pharyngeal arches. *Bioessays* **23**, 54–61 (2001).
- Quinlan, R., Martin, P. & Graham, A. The role of actin cables in directing the morphogenesis of the pharyngeal pouches. *Development* **131**, 593–599 (2004).
- Donoghue, P. C. J. & Purnell, M. A. Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.* **20**, 312–319 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions J.A.W. discovered the specimen MHNM 01-154 described herein; the other authors have contributed equally to the article.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.J. (janvier@mnhn.fr).

LETTERS

Heterotrophic plasticity and resilience in bleached corals

Andréa G. Grottoli¹, Lisa J. Rodrigues² & James E. Palardy³

Mass coral bleaching events caused by elevated seawater temperatures^{1,2} have resulted in extensive coral mortality throughout the tropics over the past few decades^{3,4}. With continued global warming, bleaching events are predicted to increase in frequency and severity, causing up to 60% coral mortality globally within the next few decades^{4–6}. Although some corals are able to recover and to survive bleaching^{7,8}, the mechanisms underlying such resilience are poorly understood. Here we show that the coral host has a significant role in recovery and resilience. Bleached and recovering *Montipora capitata* (branching) corals met more than 100% of their daily metabolic energy requirements by markedly increasing their feeding rates and CHAR (per cent contribution of heterotrophically acquired carbon to daily animal respiration), whereas *Porites compressa* (branching) and *Porites lobata* (mounding) corals did not. These findings suggest that coral species with high-CHAR capability during bleaching and recovery, irrespective of morphology, will be more resilient to bleaching events over the long term, could become the dominant coral species on reefs, and may help to safeguard affected reefs from potential local and global extinction.

Coral reefs provide essential goods and services to maritime tropical nations⁹ and are the most diverse marine ecosystems on the planet. Unfortunately, reefs are seriously declining because of global warming^{3,4}. At elevated seawater temperatures, scleractinian corals lose their endosymbiotic dinoflagellates (zooxanthellae), which renders the colony pale or white in colour, giving it a bleached appearance and often resulting in death. 'Susceptibility' of corals to bleaching has been explained primarily by coral morphology^{7,8} and zooxanthellae type or density^{10–13}. However, coral 'recovery' from bleaching events cannot be satisfactorily explained by these two factors. The role of the coral polyp itself (that is, the host) in recovery from bleaching has been studied only indirectly in a few studies^{14–16} and represents a relatively unexplored source of potential resilience in corals.

In the absence of their zooxanthellae, which can provide the coral animal with up to 100% of its daily metabolic energy (DME) requirements¹⁷, bleached and recovering corals must rely on alternative sources of fixed carbon to meet their DME needs. Stored energy reserves and heterotrophy (that is, feeding) are two such alternative sources. Whereas some coral species markedly deplete their energy reserves during bleaching^{14–16}, others do not¹⁵. However, the potential for heterotrophy as a significant source of fixed carbon for bleached and recovering corals has not been previously evaluated. Because energy reserves are a limited resource, species that can significantly increase their heterotrophic input of fixed carbon during bleaching and recovery should have an ecological advantage for long-term survival.

We considered that corals can meet their DME requirements during bleaching and recovery by consuming existing energy

reserves, by switching from acquiring fixed carbon primarily photoautotrophically to primarily heterotrophically, or by a combination of both. These hypotheses could be tested only with combined tank and field experiments. Branches from healthy *Porites compressa* and *Montipora capitata* coral colonies from Kaneohe Bay, Hawaii, were bleached in outdoor, flow-through, filtered-seawater tanks by exposing them to elevated temperatures (30 °C treatment, no zooplankton). An equal number of coral branches were kept in similar tanks at ambient seawater temperature as controls (27 °C, no zooplankton). After 30 d, half of the treatment and control branches were collected for analyses and the remaining branches were returned to the reef to recover *in situ* at ambient seawater temperatures (27 °C) and zooplankton concentrations.

In *P. compressa*, branches in the 30 °C treatment tanks bleached white and then regained some colour (pale brown) after 6 weeks of recovery. This colour change was reflected in the chlorophyll *a* (Chl *a*) concentrations (Fig. 1a) and photosynthetic rates (Fig. 1b), which decreased significantly during bleaching and had begun to recover (Fig. 1a) or had completely recovered (Fig. 1b) after 6 weeks, respectively. Total energy reserves (Fig. 1c) and total biomass (Fig. 1d) decreased significantly during bleaching, and continued to decrease during recovery. Thus, in the absence of photosynthetically and heterotrophically derived fixed carbon, *P. compressa* depleted its energy reserves significantly during bleaching. During recovery, this species continued to deplete its energy reserves, despite an increase in Chl *a* and photosynthesis and the presence of zooplankton. These findings are consistent with observations of lower Chl *a* and total lipid concentrations in bleached *P. compressa* after a natural bleaching event on the same reef in 1996 (ref. 15). Overall, these results support the hypothesis that bleached and recovering *P. compressa* corals meet their DME requirements by consuming existing energy reserves, and are largely dependant on significant inputs of zooxanthellae-derived photosynthetic carbon to recover those reserves.

In *M. capitata*, branches in the 30 °C treatment tanks bleached white and remained bleached after 6 weeks of recovery. Consistent with these observations, Chl *a* and photosynthetic rates markedly decreased during bleaching and continued to decrease during 6 weeks of recovery by a total of 97% and 90%, respectively (Fig. 1e, f). Total energy reserves (Fig. 1g) and total biomass (Fig. 1h) decreased significantly during bleaching by 39% and 34%, respectively, but were fully replenished after 6 weeks of recovery. Thus, in the absence of photosynthetically and heterotrophically derived fixed carbon, *M. capitata* significantly depleted its energy reserves and total biomass during bleaching. Once exposed to naturally available zooplankton on the reef, however, total energy reserves and total biomass were fully replenished within 6 weeks despite persistently low Chl *a* and photosynthetic rates. These results are consistent with previous findings showing that total lipids do not differ between

¹Department of Geological Sciences, Ohio State University, Columbus, Ohio 43210, USA. ²Department of Biology, Villanova University, Villanova, Pennsylvania 19085, USA.

³Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912, USA.

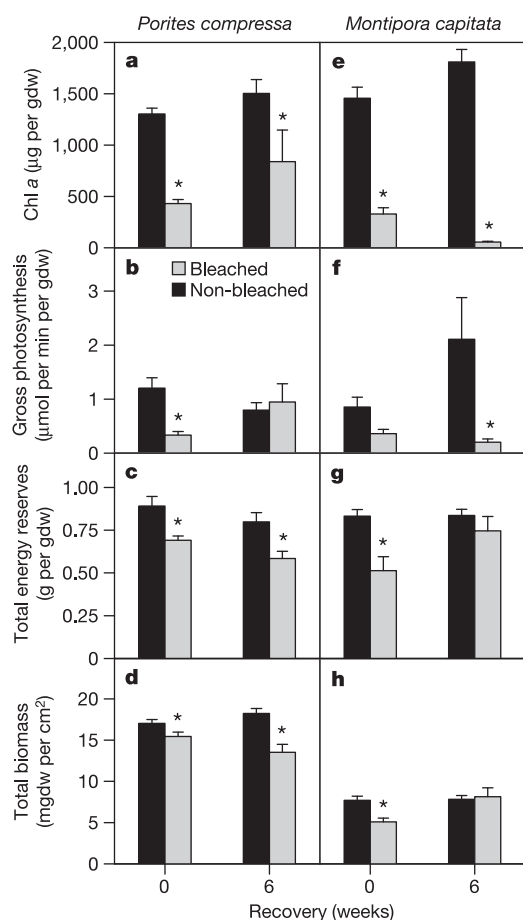


Figure 1 | Coral bleaching and recovery. Shown are Chl *a* content (a, e), gross photosynthesis (b, f), total energy reserves (c, g) and total biomass (d, h) in bleached (grey bars) and non-bleached control (black bars) *P. compressa* (a–d) and *M. capitata* (e–h) corals after 0 and 6 weeks of recovery. Data are the mean $\pm$ s.e.m. ($n = 4$ –12). Significant differences (asterisks) between averages within each recovery interval were determined by *a posteriori* tests. gdw, grams of ash-free dry weight (AFDW); mgdw, milligrams of AFDW. Additional statistical results are given in the Supplementary Information.

bleached and non-bleached corals of this species after a natural bleaching event¹⁵. Collectively, these results suggest that when zooplankton are available, bleached and recovering *M. capitata* corals may acquire large quantities of fixed carbon heterotrophically in excess of DME needs and, unlike *P. compressa*, are not dependant on symbiotic photosynthesis to recover energy reserves.

To quantify the amount of fixed carbon acquired heterotrophically in recovering *P. compressa* and *M. capitata*, a feeding experiment was conducted on bleached and non-bleached branches of each species, and on fragments of an additional species, *P. lobata* (mounding coral). For each species, feeding rates under natural conditions on the reef were determined, the daily heterotrophic carbon input (H_C) was calculated, and the CHAR (per cent contribution of heterotrophically acquired carbon to daily animal respiration (R_C)) for each coral was calculated as:

$$\text{CHAR} = \frac{H_C}{R_C} \times 100\% \quad (1)$$

CHAR is thus the percentage of a coral's DME demand that can be met through heterotrophy alone, assuming that all of the carbon in zooplankton is biologically available. These assumptions are analogous to those for CZAR (per cent contribution of zooxanthellae-acquired carbon to daily animal respiration)¹⁷, which was also calculated.

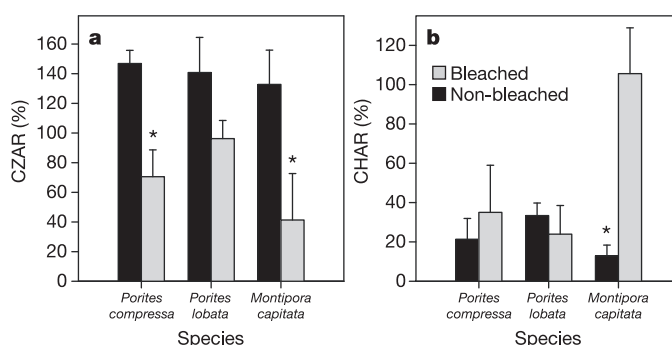


Figure 2 | Heterotrophically and photoautotrophically acquired carbon. a, Average per cent contribution of zooxanthellae-acquired carbon to daily animal respiration (CZAR). b, Average per cent contribution of heterotrophically acquired carbon to daily animal respiration (CHAR). Shown are the mean $\pm$ s.e.m. values in bleached (grey bars) and non-bleached control (black bars) *P. compressa*, *P. lobata* and *M. capitata* after 2 weeks of recovery. Significant differences (asterisks) between bleached and non-bleached corals of each species were determined by Student's *t*-test.

All three coral species had CZAR values $>100\%$ when non-bleached (controls) and $<100\%$ when bleached, decreasing by an average of 50% (Fig. 2a). Despite this reduction in photosynthetically derived carbon, feeding rates for *P. compressa* and *P. lobata* did not differ between bleached and non-bleached corals (Table 1). As a result, CHAR calculations show that for both of these species, only 21–35% of their DME demand was met heterotrophically, irrespective of bleaching status (Fig. 2b). Thus, measured decreases in the total energy reserves and biomass in bleached *P. compressa* (Fig. 1c, d) are independent of feeding rates. Therefore, with markedly diminished CZAR values *P. compressa* consumes its finite energy reserves during recovery from bleaching. Although energy reserve data are unavailable for *P. lobata* corals from Hawaii, the low-CHAR capability of bleached *P. lobata* coupled with weakened CZAR (Fig. 2a) values suggests that, like *P. compressa*, this species would also consume part or all of its energy reserves during bleaching and recovery.

In marked contrast to the two *Porites* corals, feeding rates were more than fivefold higher in bleached versus non-bleached *M. capitata* (Table 1). As a result, bleached *M. capitata* corals met their whole DME demands (average CHAR = 105%) from heterotrophy alone (Fig. 2b), more than compensating for the significantly reduced CZAR values (Fig. 2a) and enabling them to replenish their energy reserves once zooplankton was available. Non-bleached corals of this species showed the opposite pattern: low-CHAR and high-CZAR values (Fig. 2). Collectively, these findings support a previously unrecognized strategy for corals to meet metabolic demand and to maintain energy reserves and biomass, whereby some species shift from a primarily photoautotrophic to a primarily heterotrophic carbon-acquiring mode during bleaching and recovery.

The two CHAR strategies observed in bleached and recovering corals are predicted to affect coral physiology in two significant ways. First, the production of gametes (which are lipid-rich) and spawning are typically reduced or absent in corals for up to two years after a bleaching event^{4,18,19}. With depleted energy reserves and total biomass, *P. compressa* may require a prolonged recovery period before successfully spawning after bleaching. By contrast, the high energy-reserve content and total biomass of recovering *M. capitata* would facilitate uninterrupted gamete production and spawning, regardless of bleaching status. Thus, as bleaching events increase in frequency, low-CHAR corals such as *P. compressa* may not have sufficient time to replenish their energy reserves between events, resulting in lower reproductive output and selection against low-CHAR species. Second, the ability of *P. compressa* and *P. lobata* to sustain their energy demand

Table 1 | *In situ* average natural feeding rates of corals in this study

Species	Average feeding rate*		<i>t</i>	<i>P</i>
	Bleached (treatment)	Non-bleached (control)		
<i>Porites compressa</i>	10.9 (7.3)	12.8 (6.0)	0.20	>0.5
<i>Montipora capitata</i>	32.6 (4.7)	5.7 (2.3)	5.61	<0.0001
<i>Porites lobata</i>	20.6 (10.4)	18.8 (2.9)	0.17	>0.5

* The average feeding rate (± 1 s.e.m.) is the number of zooplankton caught per gram of coral AFDW per hour. Statistical results are from Student's *t*-tests for each species; feeding rates were considered significantly different at $P \leq 0.05$.

while bleached and recovering is limited by their stored energy reserves. By contrast, bleached *M. capitata* colonies should theoretically be able to sustain their DME demand indefinitely, provided zooplankton is available. During prolonged bleaching events, low-CHAR corals such as *Porites* are thus predicted to be more susceptible to mortality than high-CHAR corals such as *M. capitata*.

These findings show that energy reserves and heterotrophic capability of the coral host have a key and previously unassessed role in coral resilience to bleaching. Under future situations of increasing frequency and duration of bleaching events^{3,4,6}, we predict that coral species with high-CHAR capability will have an ecological advantage over low-CHAR species. Although a thick tissue layer or mounding morphology has been associated with recovery from past bleaching events^{7,8}, neither feature will necessarily provide sufficient energy reserves to sustain coral DME demand and reproductive output during future bleaching events and recovery. Thus, over the coming decades, coral reefs may experience a shift in coral species composition towards those with high-CHAR capability, independent of coral morphology.

METHODS

Chl *a*, photosynthesis, respiration, energy reserves and biomass. Eight branches from 12 healthy colonies of *P. compressa* and *M. capitata* (branching form) were collected at a 2-m depth in Kaneohe Bay, Hawaii (21° 26.18' N; 157° 47.56' W), in August 2003 and placed in eight outdoor flow-through seawater tanks at the Hawaii Institute of Marine Biology. The tanks were shaded with screens to simulate photosynthetically active radiation levels at a 2-m depth, and the seawater was filtered to exclude zooplankton >50 µm. On 4 September 2003, the seawater temperature was raised in four tanks by aquarium heaters (mean $\pm$ s.e.m. temperature, 30.06 $\pm$ 0.21 °C); the other four tanks were kept at ambient seawater temperature as controls (26.80 $\pm$ 0.09 °C). The corals in the 30 °C tanks gradually bleached. On 4 October 2003, photosynthesis and day and night respiration of the bleached and non-bleached control corals were measured at 30 °C and 27 °C, respectively, as described²⁰, and standardized to ash-free dry weight (AFDW). One branch per colony, per treatment and per species was collected and frozen at -80 °C. The remaining branches were placed back on the reef at a 2-m depth for 6 weeks of recovery (mean $\pm$ s.e.m. reef temperature, 26.86 $\pm$ 0.15 °C; range, 24.5–28.2 °C). On 16 November 2003, photosynthesis and respiration were measured at ambient seawater temperature, and an additional branch from each colony, species and treatment was collected and frozen at -80 °C. All remaining branches were left on the reef as part of a separate experiment.

Chlorophyll *a* (ref. 21), carbohydrates²² and proteins²³ were measured as described. Total lipid extractions were modified from ref. 15. Total energy reserves were calculated as the sum of total lipids, carbohydrates and soluble proteins per gram of AFDW. Total AFDW biomass was standardized to surface area. Four-way analysis of variance (ANOVA) was used to test for significant species, genotype, treatment and recovery interval effects in Chl *a*, photosynthesis, total energy reserves and total biomass with SAS software²⁴ (Supplementary Table 1). Because all of the coral branches were reared under the exact same conditions except temperature during the first month, differences between bleached and control corals for any of the measured variables were due to the month-long temperature treatment alone, independent of natural seasonal variation.

CHAR and CZAR. Two fragments from five healthy *P. compressa* and *M. capitata* (branching form) and *P. lobata* (mounding form) colonies were collected on 26 May 2004, bleached or not bleached (control) in tanks for 25 d, and allowed to recover on the reef *in situ* for 2 weeks as described above. Each day from 6 to 10 July 2004, a pair of bleached and control coral fragments of each species were

starved *in situ* for 8 h in isolation chambers²⁵, and then allowed to feed on ambient zooplankton for 1 h beginning 1 h after sunset. Corals were then collected and preserved in 10% formalin, and 250 polyps of each fragment were dissected as described²⁵.

Feeding rate (F_R) was calculated as the number of zooplankton caught per hour per coral fragment, standardized to coral AFDW. The proportionate contribution of individual zooplankton taxa to the coral diet (P_i) was determined on additional coral fragments as described²⁵. Average P_i did not differ between bleached and non-bleached controls of any species (multivariate analysis of variance, $P = 0.827$). The average AFDW of individual zooplankton of each taxa (M_i) and the natural abundance of each taxa were determined from nightly plankton tows. The average per cent carbon of zooplankton (C_z) in the 200–400-µm size fraction (the size corresponding to most zooplankton eaten; refs 25, 26, and J.E.P., L.J.R. and A.G.G., unpublished data) and of bulk zooplankton samples was 31% and 35%, respectively.

Daily heterotrophic carbon input (H_C) was calculated by the equation:

$$H_C = 8F_R C_z \sum_{\text{taxa}=1}^n M_i P_i \quad (2)$$

where C_z was conservatively set to 30%. H_C was calculated assuming 8 h of feeding per night²⁷, which is a conservative estimate because *P. compressa* and *M. capitata* keep their tentacles extended continuously day and night, and the contribution of microzooplankton (<50 µm) and dissolved organic matter to the coral diet, which can be significant²⁸, was not measured. Average day and night respiration rates were converted to total daily grams of carbon respired per gram of AFDW (R_C), assuming a mole-to-mole relationship of O₂ consumed to CO₂ produced during respiration. CHAR for bleached and non-bleached control corals of each species was then calculated by equation (1). The standard error of CHAR is the propagated error from the H_C and R_C calculations.

CZAR (adopted from ref. 17) was calculated by equation (3):

$$\text{CZAR} = \frac{P_C}{R_C} \times 100\% \quad (3)$$

where total daily grams of photosynthetically fixed carbon per gram of AFDW (P_C) was calculated as the sum of net photosynthetically fixed carbon plus respired carbon during the day assuming a mole-to-mole relationship of CO₂ consumed (produced) to O₂ produced (consumed) during photosynthesis (respiration).

Received 4 November 2005; accepted 3 January 2006.

- Glynn, P. Coral reef bleaching: facts, hypotheses and implications. *Global Change Biol.* **2**, 495–509 (1996).
- Jokiel, P. L. & Coles, S. L. Response of Hawaiian and other Indo-Pacific reef corals to elevated temperature. *Coral Reefs* **8**, 155–162 (1990).
- Hughes, T. P. et al. Climate change, human impacts, and the resilience of coral reefs. *Science* **301**, 929–933 (2003).
- Hoegh-Guldberg, O. Climate change, coral bleaching and the future of the world's coral reefs. *Mar. Freshwat. Res.* **50**, 839–866 (1999).
- Wilkinson, C. *Status of Coral Reefs of the World: 2000* (Australian Institute for Marine Science, Townsville, 2000).
- Buddemeier, R. W., Kleypas, J. A. & Aronson, R. B. *Coral Reefs & Global Climate Change: Potential Contributions of Climate Change to Stresses on Coral Reef Ecosystems* (Pew Center on Global Climate Change, Arlington, Virginia, 2004).
- Wilkinson, C. & Hodgson, G. Coral reefs and the 1997–1998 mass bleaching and mortality. *Nature & Resources* **35**, 16–25 (1999).
- Loya, Y. et al. Coral bleaching: the winners and the losers. *Ecol. Lett.* **4**, 122–131 (2001).
- Moberg, F. & Folke, C. Ecological goods and services of coral reef ecosystems. *Ecol. Econ.* **29**, 2151–2233 (1999).
- Glynn, P. W., Mate, J. L., Baker, A. C. & Calderon, M. O. Coral bleaching and mortality in Panama and Ecuador during the 1997–1998 El Niño Southern Oscillation event: spatial/temporal patterns and comparisons with the 1982–1983 event. *Bull. Mar. Sci.* **69**, 79–109 (2001).
- Stimson, J., Sakai, K. & Sembali, H. Interspecific comparison of the symbiotic relationship in corals with high and low rates of bleaching-induced mortality. *Coral Reefs* **21**, 409–421 (2002).
- Rowan, R. Thermal adaptation in reef coral symbionts. *Nature* **430**, 742 (2004).
- Baker, A. C. Reef corals bleach to survive change. *Nature* **411**, 765–766 (2001).
- Fitt, W. K., McFarland, F. K., Warner, M. E. & Chilcoat, G. C. Seasonal patterns of tissue biomass and densities of symbiotic dinoflagellates in reef corals and relation to coral bleaching. *Limnol. Oceanogr.* **45**, 677–685 (2000).
- Grottoli, A. G., Rodrigues, L. J. & Juarez, C. Lipids and stable carbon isotopes in two species of Hawaiian corals, *Porites compressa* and *Montipora verrucosa*, following a bleaching event. *Mar. Biol.* **145**, 621–631 (2004).

16. Porter, J. W., Fitt, W. K., Spero, H. J., Rogers, C. S. & White, M. W. Bleaching in reef corals: physiological and stable isotopic responses. *Proc. Natl Acad. Sci. USA* **86**, 9342–9346 (1989).
17. Muscatine, L., McCloskey, L. R. & Marian, R. E. Estimating the daily contribution of carbon from zooxanthellae to coral animal respiration. *Limnol. Oceanogr.* **26**, 601–611 (1981).
18. Szmant, A. M. & Gassman, N. J. The effects of prolonged 'bleaching' on the tissue biomass and reproduction of the reef coral *Montastrea annularis*. *Coral Reefs* **8**, 217–224 (1990).
19. Ward, S., Harrison, P. J. & Hoegh-Guldberg, O. Coral bleaching reduces reproduction of scleractinian corals and increases susceptibility to future stress. *Proc. 9th Int. Coral Reef Symp.* **2**, 1123–1128 (2000).
20. Lesser, M. P. Oxidative stress causes coral bleaching during exposure to elevated temperatures. *Coral Reefs* **16**, 187–192 (1997).
21. Jeffrey, S. W. & Humphrey, G. F. New spectrophotometric equations for determining chlorophylls *a*, *b*, *c1* and *c2* in higher plants, algae and natural phytoplankton. *Biochem. Physiol. Pflanz.* **167**, 191–194 (1975).
22. Dubois, M., Giles, K. A., Hamilton, J. K., Rebers, P. A. & Smith, F. Colorimetric method for determination of sugar and related substances. *Anal. Chem.* **28**, 350–356 (1956).
23. Smith, P. K. *et al.* Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150**, 76–85 (1985).
24. SAS/STAT Software Version 8.02 (SAS Institute, Cary, North Carolina, 1999–2001).
25. Palardy, J. E., Grottoli, A. G. & Matthews, K. A. Effects of upwelling, depth, morphology and polyp size on feeding in three species of Panamanian corals. *Mar. Ecol. Prog. Ser.* **300**, 79–89 (2005).
26. Palardy, J. E., Grottoli, A. G. & Matthews, K. A. Effect of naturally changing zooplankton concentrations on feeding rates of two coral species in the Eastern Pacific. *J. Exp. Mar. Biol. Ecol.* **331**, 99–107 (2006).
27. Heidelberg, K. B., Sebens, K. P. & Purcell, J. E. Composition and sources of near reef zooplankton on a Jamaican forereef along with implications for coral feeding. *Coral Reefs* **23**, 263–280 (2004).
28. Houlbrèque, F., Tambutté, E., Richard, C. & Ferrier-Pagès, C. Importance of a micro-diet for scleractinian corals. *Mar. Ecol. Prog. Ser.* **282**, 151–160 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Hawaii Institute of Marine Biology and P. Jokiel for local assistance; M. Lesser for the use of metabolism chambers and for comments; T. Pease for advice on lipid extractions; P. Petraitis for statistical advice; R. Langston for plankton collections; J. Bauer, D. Gleason and D. Alsdorf for comments; L. M. Grottoli for editorial assistance; and the University of Pennsylvania. Funding was provided by the Mellon Foundation (to A.G.G.), William Penn Fellowship (to L.J.R.), the Chemical Oceanography Program of the US National Science Foundation (to A.G.G.), and the Biological Oceanography Program of the US National Science Foundation (to A.G.G.).

Author Contributions All authors contributed equally to this work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.G.G. (grottoli.1@osu.edu).

LETTERS

Rank-related maternal effects of androgens on behaviour in wild spotted hyaenas

S. M. Dloniak¹, J. A. French² & K. E. Holekamp¹

Within any hierarchical society, an individual's social rank can have profound effects on its health and reproductive success^{1,2}, and rank-related variation in these traits is often mediated by variation in endocrine function³. Maternal effects mediated by prenatal hormone exposure are potentially important for non-genetic inheritance of phenotypic traits related to social rank³, and thus for shaping individual variation in behaviour and social structure. Here we show that androgen concentrations in wild female spotted hyaenas (*Crocuta crocuta*) are higher during late gestation in dominant females than in subordinate females. Furthermore, both male and female cubs born to mothers with high concentrations of androgens in late pregnancy exhibit higher rates of aggression and mounting behaviour than cubs born to mothers with lower androgen concentrations. Both behaviours are strongly affected in other mammals by organizational effects of androgens⁴, and both have important effects on fitness in hyaenas. Therefore, our results suggest that rank-related maternal effects of prenatal androgen exposure can adaptively influence offspring phenotype in mammals, as has previously been shown to occur in birds. They also suggest an organizational mechanism for the development of female dominance and aggressiveness in spotted hyaenas, traits that may offset the costs of extreme virilization.

Maternal effects, which occur when maternal phenotype influences offspring phenotype independent of offspring genotype, permit environmental conditions experienced by the mother to affect offspring phenotype, potentially enhancing offspring fitness⁵. In oviparous vertebrates, maternally derived androgens in eggs mediate growth and behaviour in offspring^{3,6–8}. Importantly, the social environment of a laying bird (for example, social rank and aggressive interactions) can influence the amount of androgen transferred to its eggs^{9,10}. Here we report findings from a long-term study of free-living spotted hyaenas suggesting that the social status of a pregnant female mammal can similarly affect androgen exposure of her young *in utero* and thereby modify offspring phenotype.

The society of the spotted hyaena is structured by a rigid linear dominance hierarchy, and an individual's social rank determines its priority of access to resources as well as its reproductive success^{11–13}. Within a social group, female hyaenas therefore vary considerably with respect to condition, depending on the social ranks they 'inherit' from their mothers^{12,14}. Female spotted hyaenas are generally larger than males, and they exhibit strongly masculinized genitalia through which it is uniquely difficult to give birth¹⁵. Females are also more aggressive than males¹⁶ and socially dominant to them¹¹. Large body size and aggressiveness in female hyaenas seem to be linked to genital masculinization by common endocrine or epigenetic mechanisms, although how this linkage occurs is unknown¹⁷. Except during pregnancy, female hyaenas have lower levels of circulating testosterone than do adult males¹⁸, indicating that activational effects of androgens cannot account for the sex difference in aggressiveness

documented in this species. However, the ovaries of pregnant female hyaenas produce large quantities of androgenic steroid hormones during the second half of gestation, and these cross the placenta to reach the developing fetus^{19,20}. Thus, hyaena cubs of both sexes are exposed to high concentrations of androgens during fetal life. A number of sexually dimorphic behaviours exhibited by juvenile mammals are known to be shaped by prenatal androgen exposure, and in particular, aggression and sexual play are affected by organizational effects of androgens in a wide array of mammals²¹.

Here we tested the hypothesis that prenatal exposure of cubs to

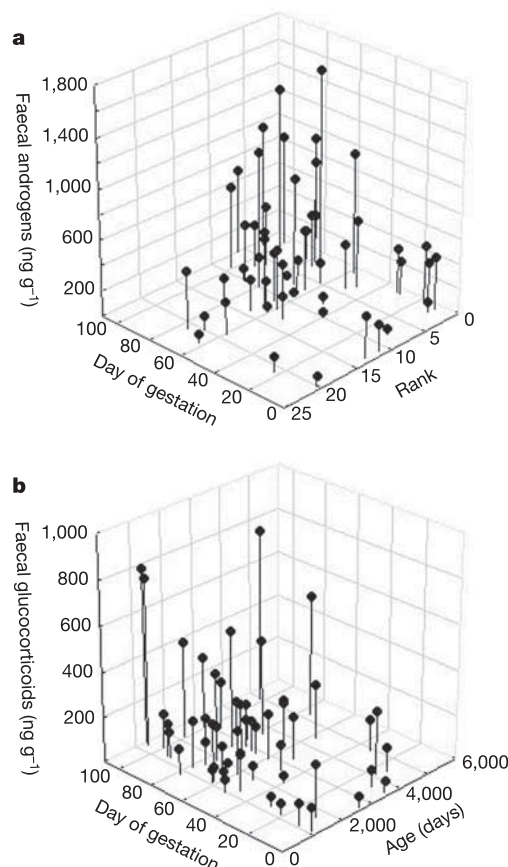


Figure 1 | Maternal hormones during gestation in wild spotted hyaenas. **a**, The graph shows plots of social rank, day of gestation and faecal androgen metabolite concentrations. **b**, The graph shows plots of maternal age, day of gestation and faecal glucocorticoid concentrations. The highest social rank possible is 1.

¹Department of Zoology, Michigan State University, East Lansing, Michigan 48824, USA. ²Departments of Psychology and Biology, University of Nebraska at Omaha, Omaha, Nebraska 68182, USA.

androgens affects offspring phenotype in hyaenas. We first investigated patterns in faecal androgen concentrations throughout gestation in pregnant females. Faecal androgen concentrations increased as gestation progressed (whole model $R = 0.61$, $F_{3,49} = 9.62$, $P = 0.004$; day of gestation partial $R = 0.45$, $P < 0.001$; Fig. 1a), and high-ranking females had higher faecal androgen concentrations than did low-ranking females during pregnancy (rank partial $R = -0.46$, $F_{1,49} = 13.26$, $P < 0.001$). Female age was not associated with variation in faecal androgen concentrations during pregnancy ($F_{1,49} = 0.71$, $P = 0.41$).

Because social rank may also influence stress in pregnant female hyaenas, and stress experienced by female mammals during pregnancy can influence offspring development via elevated glucocorticoids²², we investigated the role of faecal glucocorticoid concentrations as a covariate. The multiple regression model for faecal glucocorticoid concentrations was significant (whole model $R = 0.62$, $F_{3,49} = 10.43$, $P < 0.001$; Fig. 1b), and day of gestation explained a large proportion of the variance in faecal glucocorticoid concentrations, with faecal glucocorticoids increasing in the second half of gestation (partial $R = 0.57$, $F_{1,49} = 23.63$, $P < 0.001$). Social rank had no effect on faecal glucocorticoid concentrations in pregnant females ($F_{1,49} = 0.63$, $P = 0.43$), but older females had higher faecal glucocorticoid concentrations than younger females during gestation (partial $R = 0.38$, $F_{1,49} = 8.26$, $P = 0.006$). Thus, although several maternally associated variables influenced faecal glucocorticoid concentrations, maternal rank did not predict faecal glucocorticoid concentrations in pregnant female hyaenas.

In other mammals, day of gestation strongly affects patterns of fetal masculinization induced by androgenic hormones, with behavioural masculinization occurring later than genital masculinization²³. Because our outcome variables here were behavioural, we next inquired whether maternal rank or day of gestation affected faecal hormone concentrations in female hyaenas sampled during the second half of gestation. We also tested for relationships among litter composition, rate of maternal aggression during late pregnancy and maternal hormones by including the two former variables as covariates. Using these data, a multiple regression model of faecal androgen concentrations was still significant (whole model $R = 0.69$, $F_{5,15} = 3.08$, $P = 0.038$). High-ranking females had higher faecal androgen concentrations than low-ranking females (partial $R = -0.49$, $F_{1,15} = 6.24$, $P = 0.025$; Fig. 2), but here day of gestation was not significant ($F_{1,15} = 3.16$, $P = 0.10$). Furthermore, we found no relationship between either litter composition ($F_{2,15} = 0.22$, $P = 0.81$; Table 1) or maternal aggression rate during late pregnancy ($F_{1,15} = 0.51$, $P = 0.48$) and maternal faecal androgen concentrations. The multiple regression model for faecal glucocorticoid concentrations was significant (whole model $R = 0.83$, $F_{4,16} = 8.82$, $P < 0.001$); day of gestation influenced faecal glucocorticoid concentrations (partial $R = 0.75$, $F_{1,16} = 28.89$, $P < 0.001$), but age of

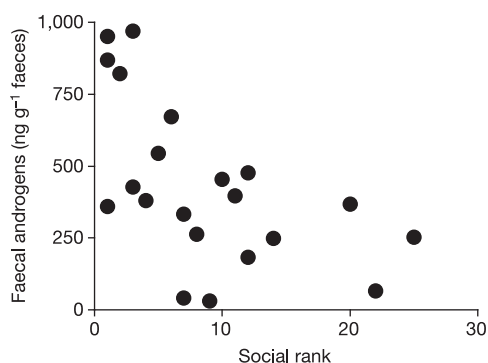


Figure 2 | The relationship between faecal androgens and social rank in pregnant female spotted hyaenas during the second half of gestation. The highest social rank possible is 1.

Table 1 | The effect of litter composition on maternal hormones in spotted hyaenas

Litter composition	Faecal androgen (ng g ⁻¹)	Faecal glucocorticoid (ng g ⁻¹)	N
Female	447.28 ± 80.97	263.66 ± 95.98	7 (4 twins, 3 single)
Mixed	429.38 ± 117.99	113.09 ± 25.60	8
Male	436.53 ± 34.42	369.28 ± 126.53	6 (2 twins, 4 single)

Values are mean ± standard error faecal androgen and faecal glucocorticoid concentrations in relation to litter composition during the second half of gestation in wild spotted hyaenas, which generally give birth to litters of one or two cubs.

mother was not significant ($F_{1,16} = 0.512$, $P = 0.48$). There was no effect of social rank ($F_{1,16} = 0.20$, $P = 0.66$) or litter composition ($F_{1,16} = 0.24$, $P = 0.63$; Table 1) on maternal faecal glucocorticoid concentrations during the second half of gestation.

Rates of aggression and mounting in cubs were positively related to maternal faecal androgen concentrations during the second half of gestation (Fig. 3). The analysis of covariance (ANCOVA) for cub mounting rate (whole model $R = 0.681$, $F_{7,27} = 6.48$, $P < 0.001$; Fig. 3a) revealed significant effects of cub sex ($F_{1,28} = 8.13$, $P = 0.008$) and maternal faecal androgen concentration ($F_{1,28} = 7.91$, $P = 0.009$), and a significant interaction between these two variables (sex × faecal androgen interaction $F_{1,28} = 17.97$, $P < 0.001$). Male cubs generally showed higher rates of play mounting than did female cubs. Maternal faecal androgen concentrations were significantly and positively related to mounting rates in both sexes, but this relationship was more robust in males than females. Neither maternal faecal glucocorticoid nor maternal rank itself significantly influenced mounting rates of cubs (faecal glucocorticoid residuals $F_{1,28} = 0.86$, $P = 0.36$; rank $F_{1,28} = 0.16$, $P = 0.69$).

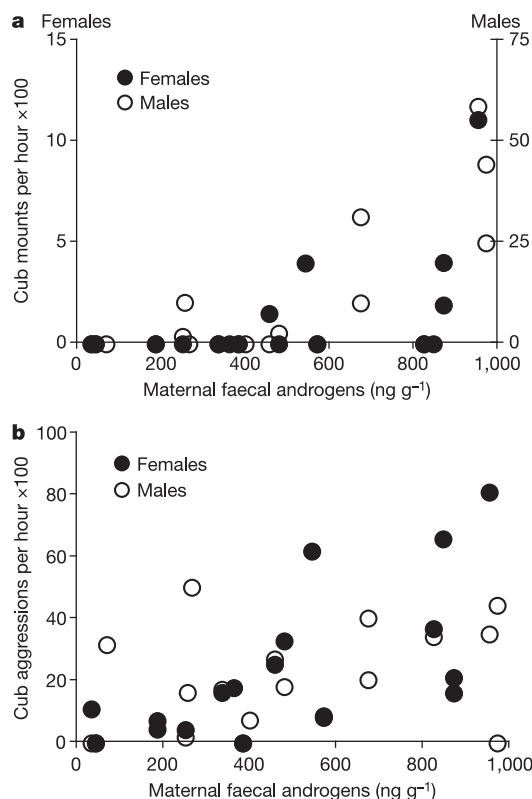


Figure 3 | Maternal androgens and cub behaviour. The relationship between maternal androgens measured during the second half of gestation and rates of mounting (a) and aggression (b) of male and female hyaena cubs aged 2–6 months. Values for mounting and aggressive behaviours are multiplied by 100 for ease of data analysis and presentation.

Similarly, we found a relationship between maternal faecal androgen concentration and cub aggression that was not directly affected by maternal rank or any other variable explored. The ANCOVA applied to the aggression data was significant (whole model $R = 0.64$, $F_{6,28} = 3.25$, $P = 0.015$; Fig. 3b). Rates of cub aggression did not vary with cub sex ($F_{1,28} = 2.745$, $P = 0.11$), maternal rank ($F_{1,28} = 0.34$, $P = 0.57$), maternal aggression during early lactation ($F_{1,28} = 2.70$, $P = 0.12$), or maternal faecal glucocorticoid residuals ($F_{1,28} = 0.99$, $P = 0.33$). However, there was a significant positive influence of maternal faecal androgen concentration on cub aggression rates ($F_{1,28} = 7.37$, $P = 0.011$), and a trend towards interaction between cub sex and maternal faecal androgen ($F_{1,28} = 8.08$, $P = 0.08$).

An extensive body of research indicates that the prenatal sex hormone environment is extremely important with respect to the morphological and behavioural development of mammalian offspring. However, to our knowledge this is the first study to demonstrate a relationship between the social rank of a free-ranging mammal and its androgen profile during pregnancy, as well as to show that this variation is related to offspring behavioural phenotype, thus suggesting a new mechanism for the transfer of status-related traits from mother to offspring via prenatal hormone exposure. Our results also offer the first definitive link between prenatal androgens and aggression in female spotted hyaenas, and they support the hypothesis that selection for aggression under intense feeding competition was the original selective pressure leading to enhanced female aggressiveness and other 'masculinized' features of female hyaenas. Aggressiveness in female spotted hyaenas should be adaptive because feeding competition in this species is very intense, and the females best able to displace conspecifics from food at ungulate kills experience the highest reproductive success^{12,13}.

Social dominance and genital masculinization in females make sex extremely challenging for male hyaenas²⁴, so males with enhanced mating ability should also be favoured by natural selection. Orienting correctly on the female during mounting is critical for the male to achieve intromission during sexual interactions, and the amount of practice a male obtains early in life may profoundly affect its lifetime reproductive success. Male hyaenas exposed to anti-androgens *in utero* experience difficulties when attempting to mate in captivity²⁵, and we expect that high-ranking sons exposed to relatively high prenatal androgen concentrations in the wild should be more successful at siring cubs than sons exposed to lower prenatal androgen concentrations. Thus, prenatal androgen exposure in both sexes may have adaptive consequences that offset the costs of female virilization in this species.

METHODS

Study population. Data were collected from hyaenas in one large, stable clan in the Masai Mara National Reserve, Kenya, from July 1993 to July 2002. This clan has been studied since 1979, and details on general methods can be found in earlier publications¹⁸. We used critical incident sampling²⁶ of all agonistic behaviour observed, and adult female social ranks were determined based on a matrix of dyadic agonistic interactions as described previously¹⁴. Age and current social rank of each female at each time of sampling were known from long-term records.

Faecal sample processing and assays. We collected 53 faecal samples from a total of 43 pregnancies of 27 adult female hyaenas. Day of gestation for a given faecal sample was determined by back-calculating from the birth date of cubs¹⁴ resulting from each pregnancy. The number of days between the sample and the birth date was subtracted from 110 days, the length of the gestation period in spotted hyaenas.

Faecal samples were collected, stored frozen and assayed for faecal androgen metabolites using an enzyme immunoassay that uses a testosterone antibody that crossreacts with dihydrotestosterone and androstenedione¹⁸. Variation in faecal androgen concentrations parallels variation in circulating testosterone in this species¹⁸. The inter-assay coefficient of variation was 7.82%, and the intra-assay coefficient of variation was 9.4%. Faecal extracts were also assayed for glucocorticoid metabolites with a modified version of an assay previously

validated for use with spotted hyaena faeces²⁷. Briefly, reconstituted faecal extracts were diluted 1:20 in steroid diluent and measured in duplicate for immunoreactive glucocorticoids by radioimmunoassay kit. This assay uses a corticosterone antibody (ICN Biomedicals) that crossreacts with the major glucocorticoid metabolites present in the faeces of spotted hyaenas²⁷. The inter-assay coefficient of variation was 11.11%, and the intra-assay coefficient of variation was 4.5%.

Behavioural data. Behavioural data were collected using critical incident sampling²⁶. Aggression rates were calculated for each adult female by dividing the total number of aggressions emitted by the total amount of time observed during the second half of gestation and the first six months of lactation. Aggression included the following behaviours: lunge, snap, bite, chase, displace, push, stand over, and intention movement to bite^{11,14}. Subject animals for behavioural analyses of offspring included all cubs born to females for which we had faecal samples from the second half of gestation, and that survived to at least 6 months of age. This included 35 cubs in eight mixed-sex twin litters, two male–male twin litters, four female–female twin litters, four singleton males and three singleton females. For each cub, we recorded all aggressions that were directed towards non-littermates, and all mounts of conspecifics. Cubs directed aggressive behaviour towards other den-dwelling cubs and older juveniles, and occasionally towards adults. Aggressive behaviour occurred in the context of competition over access to solid food as well as in other contexts. Play mounting involved one cub approaching another animal from behind, rising up on the hind legs and placing the forepaws on the other individual's back with a posture like that exhibited by adult males during copulation. For each cub, hourly rates of mounting and aggression were calculated by dividing the numbers of observed behaviours of each type by the number of hours each animal was observed during the four-month interval when cubs were 2–6 months of age. At this age, cubs have not yet attained their maternal social ranks¹⁴, so patterns of aggressive behaviour among cubs are not yet shaped by the rank relationships existing among their mothers. Cubs were observed for an average of 36.1 ± 3.8 h during this period.

Statistical analysis. To investigate the relationship between social rank and faecal steroid concentrations among pregnant females we used mixed model multiple regressions. Owing to an unbalanced design with some repeated measures, we also included individual hyaena as a random effect variable. Individual hyaena did not have a significant effect and was therefore removed from the model. We also redid these two analyses using a data set consisting of only a single random sample from each female ($N = 27$) to determine the robustness of the significant effects found in the larger data set. In these analyses, results were consistent. Faecal androgen: whole model $R = 0.62$, $F_{2,24} = 7.39$, $P = 0.003$; social rank partial $R = -0.35$, $P = 0.049$; day of gestation partial $R = 0.41$, $P = 0.02$. Faecal glucocorticoid: whole model $R = 0.70$, $F_{2,24} = 11.29$, $P < 0.001$; maternal age partial $R = 0.47$, $P = 0.016$; day of gestation partial $R = 0.62$, $P < 0.001$. We used ANCOVA to investigate variables influencing maternal hormones during the second half of gestation only, and we investigated the relationship between maternal hormones and cub behaviour with separate ANCOVAs for aggression and play mounting. Data met assumptions for the used tests and alpha was set at 0.05.

Received 17 August; accepted 19 December 2005.

- Clutton-Brock, T. H. *Reproductive Success* (Univ. Chicago Press, Chicago, 1988).
- Sapolsky, R. M. The influence of social hierarchy on primate health. *Science* **308**, 648–652 (2005).
- Schwabl, H. Yolk is a source of maternal testosterone for developing birds. *Proc. Natl Acad. Sci. USA* **90**, 11446–11450 (1993).
- Phoenix, C. H., Goy, R. W., Gerall, A. A. & Young, W. C. Organizing action of prenatally administered testosterone propionate on the tissues mediating mating behavior in the female guinea pig. *Endocrinology* **65**, 369–382 (1959).
- Mousseau, T. A. & Fox, C. W. *Maternal Effects as Adaptations* (Oxford Univ. Press, Oxford, 1998).
- Schwabl, H. Maternal testosterone in the avian egg enhances postnatal growth. *Comp. Biochem. Physiol. A* **114A**, 271–276 (1996).
- Eising, C. M., Eikenaar, C., Schwabl, H. & Groothuis, T. G. G. Maternal androgens in black-headed gull (*Larus ridibundus*) eggs: consequences for chick development. *Proc. Biol. Sci.* **268**, 839–846 (2001).
- Lipar, J. L. & Ketterson, E. D. Maternally derived yolk testosterone enhances the development of the hatching muscle in the red-winged blackbird *Agelaius phoeniceus*. *Proc. Biol. Sci.* **267**, 2005–2010 (2000).
- Müller, W., Eising, C. M., Dijkstra, C. & Groothuis, T. G. G. Sex differences in yolk hormones depend on maternal social status in Leghorn chickens (*Gallus gallus domesticus*). *Proc. Biol. Sci.* **269**, 2249–2255 (2002).
- Whittingham, L. A. & Schwabl, H. Maternal testosterone in tree swallow eggs varies with female aggression. *Anim. Behav.* **63**, 63–67 (2002).

11. Kruuk, H. *The Spotted Hyena* (Univ. Chicago Press, Chicago, 1972).
12. Frank, L. G. Social organisation of the spotted hyaena (*Crocuta crocuta*). II. Dominance and reproduction. *Anim. Behav.* **34**, 1510–1527 (1986).
13. Holekamp, K. E., Smale, L. & Szykman, M. Rank and reproduction in female spotted hyenas. *J. Reprod. Fertil.* **108**, 229–237 (1996).
14. Smale, L., Frank, L. G. & Holekamp, K. E. Ontogeny of dominance in free-living spotted hyenas: juvenile rank relations with adult females and immigrant males. *Anim. Behav.* **46**, 467–477 (1993).
15. Frank, L. G., Weldele, M. L. & Glickman, S. E. Masculinization costs in hyaenas. *Nature* **377**, 584–585 (1995).
16. Szykman, M. *et al.* Rare male aggression directed toward females in a female-dominated society: Baiting behavior in the spotted hyena. *Aggress. Behav.* **29**, 457–474 (2003).
17. Frank, L. G. in *Carnivore Behavior, Ecology, and Evolution* (ed. Gittleman, J. L.) 78–131 (Cornell Univ. Press, Ithaca, 1996).
18. Dloniak, S. M. *et al.* Non-invasive monitoring of fecal androgens in spotted hyenas (*Crocuta crocuta*). *Gen. Comp. Endocrinol.* **135**, 51–61 (2004).
19. Yalcinkaya, T. M. *et al.* A mechanism for virilization of female spotted hyenas in utero. *Science* **260**, 1929–1931 (1993).
20. Licht, P. *et al.* Androgens and masculinization of genitalia in the spotted hyaena (*Crocuta crocuta*). 1. Urogenital morphology and placental androgen production during fetal life. *J. Reprod. Fertil.* **113**, 105–116 (1998).
21. Becker, J. B. & Breedlove, S. M. in *Behavioral Endocrinology* (eds Becker, J. B., Breedlove, S. M., Crews, D. & McCarthy, M. M.) 1–38 (MIT Press, Cambridge, Massachusetts, 2002).
22. Marchlewska-Koj, A., Kapusta, J. & Kruczek, M. Prenatal stress modifies behavior in offspring of bank voles (*Clethrionomys glareolus*). *Physiol. Behav.* **79**, 671–678 (2003).
23. Goy, R. W., Bercovitch, F. B. & McBair, M. C. Behavioral masculinization is independent of genital masculinization in prenatally androgenized female rhesus macaques. *Horm. Behav.* **22**, 552–571 (1988).
24. East, M. L., Burke, T., Wilhelm, K., Greig, C. & Hofer, H. Sexual conflicts in spotted hyaenas: male and female mating tactics and their reproductive outcome with respect to age, social status, and tenure. *Proc. R. Soc. Lond. B* **270**, 1247–1254 (2003).
25. Drea, C. M. *et al.* Exposure to naturally circulating androgens during foetal life incurs direct reproductive costs in female spotted hyaenas, but is prerequisite for male mating. *Proc. R. Soc. Lond. B* **269**, 1981–1987 (2002).
26. Altmann, J. Observational study of behaviour: Sampling methods. *Behaviour* **49**, 227–256 (1974).
27. Goymann, W., Möstl, E., Van't Hof, T., East, M. L. & Hofer, H. Noninvasive fecal monitoring of glucocorticoids in spotted hyenas, *Crocuta crocuta*. *Gen. Comp. Endocrinol.* **114**, 340–348 (1999).

Acknowledgements We thank the Office of the President of Kenya for permission to conduct field research. We also thank the Kenya Wildlife Service, the Narok County Council and the Senior Warden of the Masai Mara National Reserve for their cooperation. The following individuals provided assistance in the field: E. Boydston, M. Durham, A. Engh, L. G. Frank, P. Garrett, I. Graham, T. Harty, K. Kapheim, J. Kolowski, K. Nelson, G. Ording, L. Smale, M. Szykman, J. Tanner, K. Theis, R. Van Horn, S. Wahaj, H. Watts and K. Weibel. P. Bills and P. Van Meter provided technical assistance. This research was supported by NSF grants to K.E.H. and J.A.F., and fellowships from Michigan State University and the International Foundation for Ethical Research to S.M.D.

Author Contributions Laboratory work, including development and validation of faecal hormone assays, was performed by S.M.D. and J.A.F. Field work was conducted by S.M.D. and K.E.H. The original idea for the study was conceived by K.E.H. S.M.D. conducted the statistical analyses and drafted the manuscript, which was further developed by all three authors.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.M.D. (dloniak@kenyaweb.com).

LETTERS

The DNA sequence, annotation and analysis of human chromosome 3

Donna M. Muzny¹, Steven E. Scherer¹, Rajinder Kaul², Jing Wang³, Jun Yu³, Ralf Sudbrak^{4,5}, Christian J. Buhay¹, Rui Chen¹, Andrew Cree¹, Yan Ding¹, Shannon Dugan-Rocha¹, Rachel Gill¹, Preethi Gunaratne¹, R. Alan Harris¹, Alicia C. Hawes¹, Judith Hernandez¹, Anne V. Hodgson¹, Jennifer Hume¹, Andrew Jackson¹, Ziad Mohid Khan¹, Christie Kovar-Smith¹, Lora R. Lewis¹, Ryan J. Lozado¹, Michael L. Metzker¹, Aleksandar Milosavljevic¹, George R. Miner¹, Margaret B. Morgan¹, Lynne V. Nazareth¹, Graham Scott¹, Erica Sodergren¹, Xing-Zhi Song¹, David Steffen¹, Sharon Wei¹, David A. Wheeler¹, Mathew W. Wright⁶, Kim C. Worley¹, Ye Yuan¹, Zhengdong Zhang¹, Charles Q. Adams¹, M. Ali Ansari-Lari¹, Mulu Ayele¹, Mary J. Brown¹, Guan Chen¹, Zhijian Chen¹, James Clendenning², Kerstin P. Clerc-Blankenburg¹, Runsheng Chen³, Zhu Chen³, Clay Davis¹, Oliver Delgado¹, Huyen H. Dinh¹, Wei Dong³, Heather Draper¹, Stephen Ernst², Gang Fu³, Manuel L. Gonzalez-Garay¹, Dawn K. Garcia⁷, Will Gillett², Jun Gu³, Bailin Hao³, Eric Haugen², Paul Havlak¹, Xin He⁷, Steffen Hennig⁸, Songnian Hu³, Wei Huang³, Laronda R. Jackson¹, Leni S. Jacob¹, Susan H. Kelly¹, Michael Kube⁴, Ruth Levy², Zhangwan Li¹, Bin Liu³, Jing Liu¹, Wen Liu¹, Jing Lu¹, Manjula Maheshwari¹, Bao-Viet Nguyen¹, Geoffrey O. Okwuonu¹, Anthony Palmeiri², Shiran Pasternak¹, Lesette M. Perez¹, Karen A. Phelps², Farah J. H. Plopper¹, Boqin Qiang³, Christopher Raymond², Ruben Rodriguez⁷, Channakhone Saenphimmachak², Jireh Santibanez¹, Hua Shen¹, Yan Shen³, Sandhya Subramanian², Paul E. Tabor¹, Daniel Verduzco¹, Lenee Waldron¹, Jian Wang³, Jun Wang³, Qiaoyan Wang¹, Gabrielle A. Williams¹, Gane K.-S. Wong³, Zhijian Yao³, JingKun Zhang¹, Xiuqing Zhang³, Guoping Zhao³, Jianling Zhou¹, Yang Zhou², further contributors*, David Nelson¹, Hans Lehrach⁴, Richard Reinhardt⁴, Susan L. Naylor⁷, Huanming Yang³, Maynard Olson², George Weinstock¹ & Richard A. Gibbs¹

After the completion of a draft human genome sequence¹, the International Human Genome Sequencing Consortium has proceeded to finish² and annotate each of the 24 chromosomes comprising the human genome. Here we describe the sequencing and analysis of human chromosome 3, one of the largest human chromosomes. Chromosome 3 comprises just four contigs, one of which currently represents the longest unbroken stretch of finished DNA sequence known so far. The chromosome is remarkable in having the lowest rate of segmental duplication in the genome. It also includes a chemokine receptor gene cluster as well as numerous loci involved in multiple human cancers such as the gene encoding FHIT, which contains the most common constitutive fragile site in the genome, *FRA3B*³. Using genomic sequence from chimpanzee and rhesus macaque, we were able to characterize the breakpoints defining a large pericentric inversion that occurred some time after the split of Homininae from Ponginae, and propose an evolutionary history of the inversion.

The physical map of chromosome 3 was generated using a combination of STS-derived probe screening of bacterial artificial chromosome (BAC) clone libraries and the fingerprint map⁴ and then used to pick concomitantly a tiling path of 1,710 overlapping BAC and P1-derived artificial chromosome (PAC) clones for

sequencing. The two remaining euchromatic gaps have proved recalcitrant to screening BAC libraries (<http://bacpac.chori.org/>) consisting of better than an 80-fold representation of the human genome. Gap sizes were estimated by a combination of fibre-fluorescence *in situ* hybridization (FISH; C. Wagner-McPherson, personal communication) and homologous gap flank mapping to the chimpanzee and/or rhesus macaque assembly, and total an estimated 137 kilobases (kb). A more thorough cross-species analysis of gap size can be found in Supplementary Table 1. The data extend to within 35 kb of the (TTAGGG)_n telomeric repeat motif on the p-arm and 55 kb on the q-arm of chromosome 3 (H. Riethman, personal communication; see also <http://www.wistar.upenn.edu/Riethman/>). The p-arm pericentromeric sequence contains 147.5 kb of monomeric alpha-satellite repeats, whereas the q-arm sequence extends 8.6 kb into these repeats. The chromosome is characterized by a highly polymorphic heterochromatin block at 3q11.2—similar to, but far shorter than, those present on chromosomes 1, 9, 16 and Y—that ranges in size from 0.2 to 2.0 megabases (Mb)⁵ and is thought to consist primarily of satellite 1 repeat sequence⁶. We have assumed a 1.5-Mb block and a core centromere size of 2.9 Mb to arrive at an overall chromosome length of 199,344,050 base pairs (bp).

The sequence was generated using a clone-by-clone random

¹Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ²Department of Medicine, Division of Medical Genetics, University of Washington Genome Center, Fluke Hall on Mason Rd, Box 352145 Seattle, Washington 98195, USA. ³Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China; James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310008, China; Chinese National Human Genome Center at Shanghai (CHGC), Shanghai 201203, China; and Chinese National Human Genome Center, Beijing (CHGB), Beijing 100176, China. ⁴Max Planck Institute for Molecular Genetics, 14195 Berlin-Dahlem, Germany. ⁵Institute for Clinical Molecular Biology, Christian-Albrechts University, 24105 Kiel, Germany. ⁶HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK. ⁷University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78229, USA. ⁸RZPD German Resource Center for Genome Research, 14059 Berlin, Germany.

*Lists of further contributors and their affiliations appear in the Supplementary Information.

shotgun sequencing and finishing strategy² (see Methods). Each tiling path BAC clone was finished to community standards (<http://genomeold.wustl.edu/Overview/g16stand.php>). We finished 194,944,050 bp of euchromatic sequence to an independently measured accuracy of at least 99.99%⁷ and have covered more than 99.99% of the euchromatic chromosome. Each of the landscape features and annotations outlined here may be viewed as user-specified tracks on the Genboree Browser (<http://www.genboree.org/Hs.chr3>).

In the current assembly of the genome (NCBI build 35) all RefSeq⁸ genes are entirely accounted for with partial sequence available for *SLC25A26* (NM_173471 and splice variants; bases encoding the first exon are now accounted for in GenBank, accession AC170165) and *RYBP* (NM_012234; bases 1–218 may be polymorphic in the population). As can be seen in Supplementary Fig. 1, there is strong concordance in marker order and content between the finished sequence and various genetic maps (see Supplementary Methods).

We analysed the recombination rate across the chromosome using the deCODE⁹ markers and found the statistics to be in line with the other human chromosomes, yielding a sex-average rate of 1.14 cM Mb⁻¹. The female and male recombination rates were found to be 1.43 cM Mb⁻¹ and 0.85 cM Mb⁻¹ respectively, with maximum rates of 3.77 cM Mb⁻¹ in females and 5.77 cM Mb⁻¹ in males (Supplementary Fig. 2). Although there are no recombination deserts as previously defined¹⁰, there is a recombination jungle at the tip of the p-arm (3p26.3–26.1).

As the beginning of what is inherently a dynamic process, we used manual curation of the automated Ensembl annotation output of NCBI Human Assembly build 33 to characterize fully the gene content of chromosome 3. Using all publicly available human protein, complementary DNA and spliced expressed sequence tag (EST) databases together with selected gene prediction algorithms and UCSC cDNA resources, we characterized each locus using the standards established by the Human Annotation Working Group

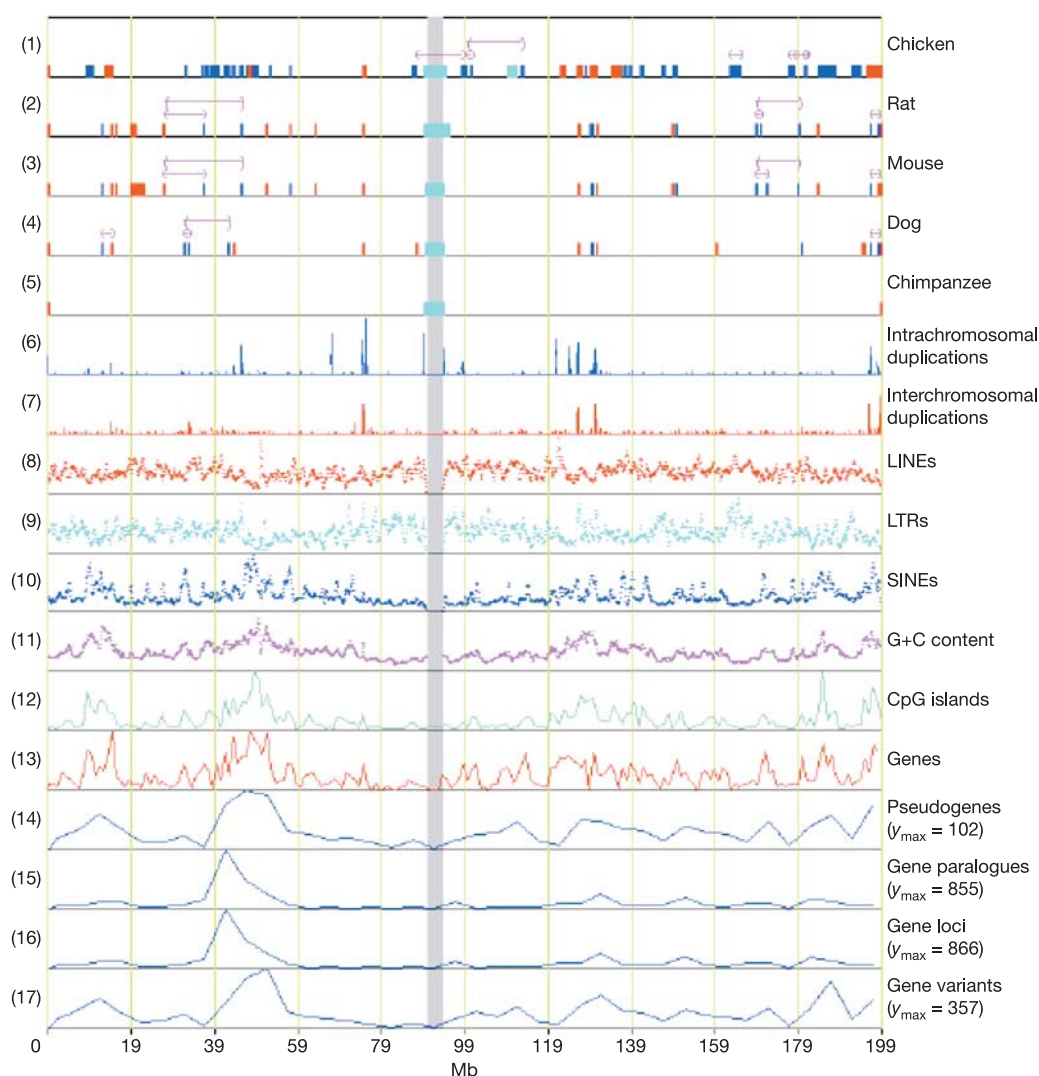


Figure 1 | Correlation of syntenic breakpoints with general chromosome landscape features. Tracks are numbered on the left and syntenic alignments across the human chromosome are shown in the top five tracks: (1) human–chicken; (2) human–rat; (3) human–mouse; (4) human–dog; and (5) human–chimpanzee. The inter- and intrachromosomal breakpoints are represented by red and blue gaps, respectively. Cyan gaps indicate regions without sequence alignment and the centromere is located in the cyan gap that is common to all species. Purple brackets indicate sequence inversions. The density of recent segmental intra- and interchromosomal

duplications from low-copy repeats is shown in tracks (6) and (7). The incidence of major interspersed (high-copy) repeats are depicted in tracks (8), (9) and (10) for LINEs, LTRs and SINEs, respectively. The variations in G+C content, and densities of CpG islands, genes and pseudogenes appear in tracks (11), (12), (13) and (14), respectively, whereas gene paralogue density, gene density and gene variant density appear in tracks (15), (16) and (17), respectively. Gene density in track 13 is from UCSC ‘known genes’, whereas track 16 reflects the non-redundant locus annotations detailed in this study.

(<http://www.sanger.ac.uk/HGP/havana/hawk.shtml>). Starting with 1,249 loci and 1,697 variants, we annotated 1,585 gene loci (Fig. 1). Among these were 1,425 known coding genes, 8 novel genes, 27 novel transcripts, 3 putative genes and 122 pseudogenes. We found 4,857 paralogous gene pairs, just 361 of which were intrachromosomal—a further reflection of the low segmental duplication (greater than 90% similarity over at least 1 kb) rate on this chromosome. However, this paralogous set reflects a number of ancient duplications, including one that contains genes encoding the developmentally important nuclear receptors (see Supplementary Fig. 3). Excluding the pseudogenes, the average gene density is 8.8 genes per Mb, making it one of the more gene-poor chromosomes. However, although the average gene density is low, as with other gene-poor chromosomes such as chromosome 13, the genes are larger than the genome average and cover 98.3 Mb or 49% of the chromosome. Chromosome 3 contains two gene-dense clusters on the p-arm between base coordinates 10–17 Mb and 41–55 Mb (18.9 and 21.1 genes per Mb, respectively). These two regions alone account for 26% of the genes on the chromosome. Relatively gene-poor tracts are confined to the pericentromeric regions.

Approximately 57% of chromosome 3 genes expressed alternative transcripts with an average of 2.86 transcripts per gene. The *IFRD2* gene had the highest number of alternative transcripts at 21 annotated variants. Most of these could produce altered protein products (3,163 different proteins from among 4,096 alternative transcripts). There were at least 681 partial transcripts in the database for which we could not identify the complete coding sequence.

We analysed the chromosome for the presence of distinguishing features including CpG islands, G+C content, segmental duplications, repeat content (see Fig. 1) and non-coding RNA. Of the 1,575 genetic loci analysed (including their variants), 56–57% were associated with a CpG island. The G+C content was found to correlate well with gene density, as expected, and repeat content is

unremarkable. Chromosome 3 is relatively devoid of segmental duplications, having just 1.7% of its bases composed of duplicated sequence compared to the whole-genome average of 5.3%. This is the lowest percentage for any chromosome in the genome.

We analysed the known and predicted non-coding RNA gene content on chromosome 3 as a prelude to future annotation of regulatory regions. Using three different strategies (see Methods), we were able to find 703 putative, non-redundant non-coding RNAs (ncRNAs). The most abundant ncRNA candidates found (68%) were mRNA-like ncRNAs, whereas the remainder consisted of smaller ncRNAs of various types including small nuclear RNAs, Y RNAs, small nucleolar RNAs, microRNAs, SRP RNAs, a telomerase RNA, 7SK RNAs, small Cajal body-specific RNAs (scaRNAs), a small non-messenger RNA (snmRNA) and a small group of ribosomal RNAs and transfer RNAs (see Supplementary Tables 2 and 3). Further characterization of the genomic landscape from 3pter to D3S3397 is also available¹¹.

Cytogenetic studies using chromosome painting and comparative mapping analysis suggest that a fission event in the largest ancestral eutherian chromosome gave rise to human chromosomes 3 and 21 (ref. 12). These observations were extended by a study¹³ using gene or genome sequence anchors and the chicken genome sequence as an out-group to reconstruct the ancestral mammalian genome. These analyses paint a more complicated evolutionary picture requiring six or seven recombination events to account for human chromosome 3. The pattern gets more complicated in comparison to the rodent genomes due to their well-characterized higher rates of interchromosomal rearrangement. Nevertheless, consistent syntenic blocks are observed in both mouse and rat, particularly at each end of the chromosome and along most of the q-arm (see Fig. 1).

Further comparative FISH analysis revealed that a large-scale pericentric inversion occurred in the ancestor of the African apes and is present in modern human chromosome 3 as well as the

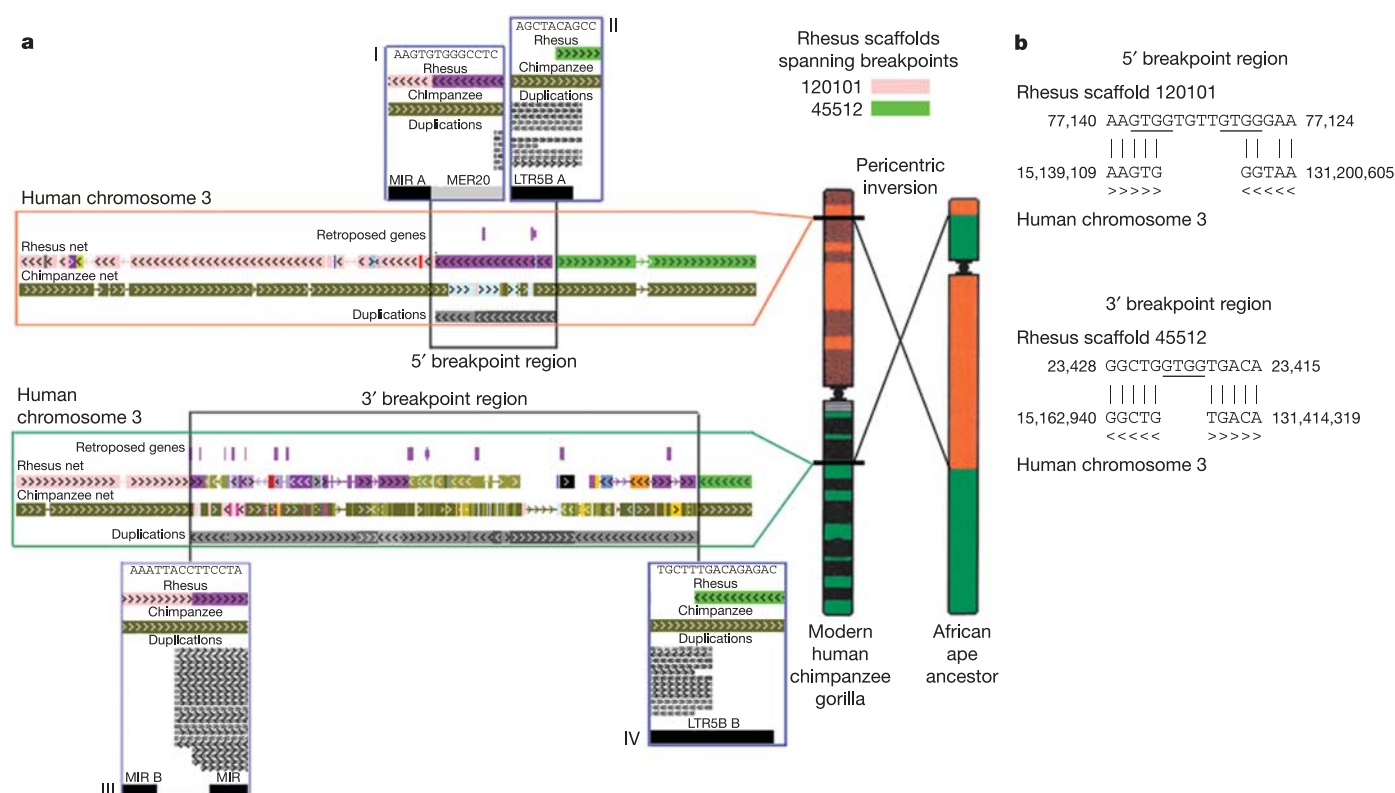


Figure 2 | Human chromosome 3 pericentric inversion breakpoints. **a**, Inversion breakpoint regions in human chromosome 3 compared to chimpanzee and rhesus macaque. Insets of breakpoint boundaries are

designed by Roman numerals. Figure adapted from the UCSC genome browser and ref. 16. **b**, Rhesus breakpoint sequence aligned to human. The sequences homologous to both breakpoints are underlined.

chimpanzee and gorilla orthologues, but not in orang-utan or Old World monkeys¹⁴. Two scaffolds from the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) rhesus macaque Mmul_0.1 assembly were found to span both breakpoints of the human inversion (Fig. 2; see Supplementary Table 4 for breakpoint details). The macaque 5' breakpoint is characterized by a short homologous GTGG track (Fig. 2b) and by a mammalian interspersed repeat (MIR) that was split by a segmental duplication before the inversion resulting in one part, designated MIR A, present in boundary I and a second part, designated MIR B, present in boundary III (see Fig. 2a). The MIR at the 3' end of boundary III was present in the segmental duplication and may have been involved in the insertion event. A number of simple repeats and low complexity regions were found within 1 kb of the breakpoint (see Supplementary Table 5). Each of these elements, including retrotransposons¹⁵, short homologous sequence and alternating purine-pyrimidine tracks¹⁶ have been reported for many other breakpoints.

The inversion breakpoint regions on human chromosome 3 are characterized by segmental duplications. Both breakpoints contain segmental duplications that are at least partially repeated at numerous intra- and interchromosomal locations on chromosomes 3, 4, 7, 8, 11, 12 and 16. The entire 5' segmental duplication maps to a location on 11q13, suggesting its probable origin, and the 3' segmental duplication maps to an adjacent block of 11q13. The 5' and 3' segmental duplications do not align to one another. These results suggest that a single segmental duplication occurred at the 5' breakpoint followed by the inversion break within the segmental duplication, splitting a long terminal repeat from the human endogenous retrovirus K family (LTR5B) into two parts. The two parts are designated LTR5B A and LTR5B B in Fig. 2a, and the dot plots in Supplementary Fig. 4 can be placed together to form a complete LTR5B element. The 5' and 3' segmental duplications are aligned to the same chromosome 11 region in this figure, showing their adjacency and also that the duplication inserted in the reverse orientation on chromosome 3 compared with chromosome 11.

It is unclear whether segmental duplications are the cause or result of rearrangements¹⁵. The segmental duplication is not present in macaque, and the MIR element spanning the 5' breakpoint in macaque seems to have been split by the segmental duplication before the inversion. LTR5B elements are found in human, chimpanzee and gorilla but not orang-utan¹⁷, suggesting that the duplication occurred after the African ape-orang-utan divergence. Indeed, the 3' duplication is not present in orang-utan or gibbon based on comparative FISH studies¹⁸. The LTR5B element was present in the segmental duplication, so the splitting of LTR5B by the inversion most probably occurred after the duplication.

Regions of segmental duplications involved in evolutionary rearrangements can also be involved in rearrangements associated with human disease^{19,20}. The q-arm pericentromeric breakpoint undergoes t(3;11)(q21;q13) translocations in head and neck squamous cell carcinomas²¹ and acute myeloid leukaemia²². Perhaps the most interesting—because it involves the same regions as the evolutionary inversion—are inv(3)(p25;q21) pericentric inversions, along with other accompanying chromosomal abnormalities, which cause severe developmental abnormalities^{23,24}.

At least 505 disease loci have been mapped to chromosome 3 (see <http://www.ncbi.nlm.nih.gov/Omim/mimstats.html> and Supplementary Tables 6 and 7). These include simple repeat expansions such as spinocerebellar ataxia 7 (*ATXN7*) involving an expanded CAG repeat (38–150 copies in the mutant allele compared with 7–17 copies in the normal allele) and myotonic dystrophy 2 (*DM2*), caused by the expansion of a CCTG repeat in the zinc finger gene *ZNF9*. Genes involved in DNA repair are encoded on chromosome 3, including *XPC* (xeroderma pigmentosum complementation group C), *MLH1*, a gene involved in DNA mismatch repair and mutated in hereditary non-polyposis coli, and Fanconi anaemia complementation group D2 (*FANCD2*), mapped to 3p26.

Among the most interesting medically relevant regions on the chromosome is the cluster of chemokine receptor genes mapping to 3p21. Within this group, the gene encoding CCR5 has been shown to be a critical cofactor for HIV-1 virus entry into cells, as defective alleles have been associated with HIV infection resistance. The clustering of both the chemokine and chemokine receptor genes suggests a relatively recent and rapid evolution of both gene families by local duplications.

Finally, a large number of cancer lesions have been mapped to chromosome 3 and cancer breakpoints seem to correlate with the four known fragile sites on the chromosome (see Supplementary Fig. 5). The cancer loci include the *VHL* gene, mutated in von Hippel–Lindau syndrome and linked to kidney cancer susceptibility, β -catenin, mutated in a number of colon tumours, and mutations in the *FHIT* gene, which encompasses the most common fragile site in the human genome (*FRA3B*) and for which aberrant transcripts have been found in about half of all oesophageal, stomach and colon carcinomas. The complete chromosome 3 sequence presented here provides a rich resource for future studies aimed at understanding our evolutionary history and the molecular basis of human variation and disease.

METHODS

Mapping and sequencing. BAC clone screening, sequencing and finishing strategies are described in Supplementary Methods. Sequence overlaps between BAC clones were verified by BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>) and polymorphic regions within overlaps were confirmed by polymerase chain reaction using a bi-gender, multi-ethnic pool of genomic DNA isolated from eight individuals (J. Belmont, personal communication). The quality of the assembly was assessed using restriction digests of individual BAC clones together with genetic and radiation hybrid map marker content and order (see Supplementary Methods and Supplementary Fig. 1), gene content and large-insert paired ends. Unique fosmid end sequences (Broad Institute) were downloaded from the UCSC Genome Browser, aligned to the genomic sequence and checked for both pair orientation and resulting insert size.

Annotation. We manually curated each known, novel CDS and novel transcript locus, defined as a set of one or more transcripts that share at least one exon of in-frame coding sequence and supported by full-length and partial human cDNAs or vertebrate cDNAs having a best in genome BLAST/Blat (<http://genome.ucsc.edu/cgi-bin/hgBlat>) hit with >98% identity. The cDNAs/reference sequences (RefSeqs)⁸ were compared to the genomic sequence to place exons, and all splice sites were examined for canonical sequence. Coding regions were examined for a best-fit open reading frame. The 5' and 3' untranslated regions were annotated and extended using available EST and cDNA evidence; poly-A sites and poly-A signals were annotated on each gene where identified. Alternative splice variants were identified from cDNA, EST and protein evidence and the translation product for each CDS was verified using SwissProt. Pseudogenes were defined as sequences with no direct evidence for expression while having a match with high score to a spliced mRNA or spliced EST from elsewhere in the genome. This is a more stringent definition than has been applied by others in broad genomic screens of pseudogenes and results in a fivefold lower count across chromosome 3 than previously reported²⁵. For paralogue analysis, protein sequences corresponding to the 'KnownGenes' track of the UCSC Browser were compared in an all-against-all BLAST search. Two loci were defined as paralogues if there was a match of any of their transcript variants with the following criteria: expect value cutoff of 10^{-10} or less, the lengths of the matching transcripts are within 20% of each other, and the match length extends over 70% of the average length of the two sequences. The complete set of annotations has been submitted to the Vega database (http://vega.sanger.ac.uk/Homo_Sapiens/).

Landscape features. CpG islands were defined as an expanse of greater than 200 nucleotides in which the G+C content is >50% and the ratio of the observed CG dinucleotides to expected in the segment is >0.6. We scanned the chromosome for ncRNAs as detailed in Supplementary Methods. We identified recent intra- and interchromosomal segmental duplications by using BLAST to align the repeat-masked chromosome sequence against itself and the rest of the human genome. The duplication densities were calculated by averaging the duplications of each base over non-overlapping 100-kb windows after filtering low-identity matches (<90%). The densities of short interspersed elements (SINEs), long interspersed elements (LINEs) and long terminal repeats (LTRs) were calculated from repeat-masked data using 100-kb windows. The G+C density was calculated by counting the G+C content over non-overlapping 100-kb windows.

The densities of CpG islands, genes (BCM-HGSC annotations) and pseudogenes were counted and displayed using 1-Mb windows.

Comparative analysis. The multiple alignments of human, chimpanzee (panTro1), dog (canFam1), mouse (mm5), rat (rn3), chicken (galGal2), zebrafish (danRer1) and *Fugu* (fr1) were downloaded from the UCSC Genome Browser (<http://hgdownload.cse.ucsc.edu/>) (see Supplementary Methods). The pairwise synteny blocks between human and other species were parsed with Synteny-Parser (X. Song and G. Weinstock, unpublished perl script), which was tuned to include all visible chromosome rearrangements in the dot plot. Rhesus scaffolds from the Mmul_0.1 preliminary assembly were mapped to human chromosome 3 using Pash²⁶. Rhesus scaffolds mapped by both Pash and human-rhesus net alignments (UCSC) were aligned with orthologous human regions and chimpanzee regions from the human-chimpanzee reciprocal best chain alignments (UCSC) using MLAGAN²⁷.

Received 24 October 2005; accepted 17 March 2006.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
3. Markkanen, A., Heinonen, K., Knuutila, S. & de la Chapelle, A. Methotrexate-induced increase in gap formation in human chromosome band 3p14. *Hereditas* **96**, 317–319 (1982).
4. McPherson, J. D. *et al.* A physical map of the human genome. *Nature* **409**, 934–941 (2001).
5. Kalitsis, P., Earle, E., Vissel, B., Shaffer, L. G. & Choo, K. H. A chromosome 13-specific human satellite I DNA subfamily with minor presence on chromosome 21: further studies on Robertsonian translocations. *Genomics* **16**, 104–112 (1993).
6. Tagarro, I., Fernandez-Peralta, A. M. & Gonzalez-Aguilera, J. J. Digestion of centromeric DNA from each human metaphase chromosome by the 6 bp-restriction enzyme *StuI*. *Histochemistry* **99**, 453–456 (1993).
7. Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429**, 365–368 (2004).
8. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
9. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
10. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
11. The Chinese Human Genome Sequencing Consortium. “Beijing Region” (3pter-D3S3397) of the human genome: complete sequence and analysis. *Sci. China C Life Sci.* **48**, 311–329 (2005).
12. Wienberg, J. The evolution of eutherian chromosomes. *Curr. Opin. Genet. Dev.* **14**, 657–666 (2004).
13. Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A. & Tesler, G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.* **15**, 98–110 (2005).
14. Ventura, M. *et al.* Recurrent sites for new centromere seeding. *Genome Res.* **14**, 1696–1703 (2004).
15. Eichler, E. E. & Sankoff, D. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**, 793–797 (2003).
16. Bacolla, A. *et al.* Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl Acad. Sci. USA* **101**, 14162–14167 (2004).
17. Hughes, J. F. & Coffin, J. M. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nature Genet.* **29**, 487–489 (2001).
18. Yue, Y., Grossmann, B., Ferguson-Smith, M., Yang, F. & Haaf, T. Comparative cytogenetics of human chromosome 3q21.3 reveals a hot spot for ectopic recombination in hominoid evolution. *Genomics* **85**, 36–47 (2005).
19. Eichler, E. E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
20. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
21. Jin, Y., Jin, C., Wennerberg, J., Hoglund, M. & Mertens, F. Cyclin D1 amplification in chromosomal band 11q13 is associated with overrepresentation of 3q21-q29 in head and neck carcinomas. *Int. J. Cancer* **98**, 475–479 (2002).
22. Cigudosa, J. C. *et al.* A recurrent translocation, t(3;11)(q21;q13), found in two distinct cases of acute myeloid leukemia. *Cancer Genet. Cytogenet.* **83**, 119–120 (1995).
23. Allderdice, P. W., Browne, N. & Murphy, D. P. Chromosome 3 duplication q21 leads to qter deletion p25 leads to pter syndrome in children of carriers of a pericentric inversion inv(3) (p25q21). *Am. J. Hum. Genet.* **27**, 699–718 (1975).
24. Stine, S. B., Clark, C. E., Telfer, M. A., Casey, P. A. & Cowell, H. R. Ullrich-Turner syndrome (45,X/46,X,i[Xq]) in a child with a familial inversion of chromosome 3. *Am. J. Med. Genet.* **12**, 57–62 (1982).
25. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13**, 2541–2558 (2003).
26. Kalafus, K. J., Jackson, A. R. & Milosavljevic, A. Pash: efficient genome-scale sequence anchoring by Positional Hashing. *Genome Res.* **14**, 672–678 (2004).
27. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We wish to acknowledge T. Taylor and the rest of the Riken Genome Center for substitute BAC clones added to the tiling path. The authors acknowledge and thank the genome sequencing community for generating the data sets used in our comparative analysis. We also acknowledge the following authors from the HUGO Gene Nomenclature Committee: S. Povey (chair), E. A. Bruford, V. K. Khodiyar, R. C. Lovering, M. J. Lush, K. M. B. Sneddon, T. P. Sneddon and C. C. Talbot Jr. This work was supported by NIH grants to M. Olson and R. Gibbs. The Chinese Human Genome Sequencing Consortium is sponsored by the Ministry of Science and Technology, Chinese Academy of Sciences, National Natural Science Foundation of China, Beijing Municipal Government, Zhejiang Provincial Government, Hangzhou Municipal Government and Yueqing Municipal Government. Funding was also supplied by the Federal German Ministry of Education and Research and The Max Planck Society.

Author Information The chromosome 3 sequence has been deposited in GenBank under accession number NC_000003. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.S. (sscherer@bcm.tmc.edu) or H.Y. (yanghm@genomics.org.cn).

Pluripotency of spermatogonial stem cells from adult mouse testis

Kaomei Guan^{1*}, Karim Nayernia^{2*}, Lars S. Maier¹, Stefan Wagner¹, Ralf Dressel³, Jae Ho Lee², Jessica Nolte², Frieder Wolf¹, Manyu Li², Wolfgang Engel² & Gerd Hasenfuss¹

Embryonic germ cells as well as germline stem cells from neonatal mouse testis are pluripotent and have differentiation potential similar to embryonic stem cells^{1,2}, suggesting that the germline lineage may retain the ability to generate pluripotent cells. However, until now there has been no evidence for the pluripotency and plasticity of adult spermatogonial stem cells (SSCs), which are responsible for maintaining spermatogenesis throughout life in the male³. Here we show the isolation of SSCs from adult mouse testis using genetic selection, with a success rate of 27%. These isolated SSCs respond to culture conditions and acquire embryonic stem cell properties. We name these cells multipotent adult germline stem cells (maGSCs). They are able to spontaneously differentiate into derivatives of the three embryonic germ layers *in vitro* and generate teratomas in immunodeficient mice. When injected into an early blastocyst, SSCs contribute to the development of various organs and show germline transmission. Thus, the capacity to form multipotent cells persists in adult mouse testis. Establishment of human maGSCs from testicular biopsies may allow individual cell-based therapy without the ethical and immunological problems associated with human embryonic stem cells. Furthermore, these cells may provide new opportunities to study genetic diseases in various cell lineages.

For the isolation of spermatogonial stem cells, we used the spermatogonia-specific marker *Stra8* (ref. 4). The activity of its regulatory sequences enables the enrichment of germline stem cells from transgenic mice^{5,6}. Testicular cells ($2\text{--}3 \times 10^7$ per mouse) were isolated from *Stra8*-enhanced green fluorescent protein (EGFP)/*Rosa26* adult mice ($n = 15$; Supplementary Fig. 1) and cultured in medium containing glial-derived neurotrophic factor (GDNF) for 4–7 days. Thereafter, GFP⁺ cells ($1\text{--}2 \times 10^6$ cells) were isolated using fluorescence-activated cell sorting (FACS). To prove the stem-cell activity of sorted GFP⁺ cells, we transplanted these cells into seminiferous tubuli of germ-cell-depleted mice. After 5–6 months, regeneration of normal spermatogenesis was observed (data not shown). Testicular teratomas (tumours derived from pluripotent cells) are rare in most inbred mouse strains, and occur spontaneously at a rate of 1–5% in 129/Sv inbred strains⁷. In a mixed C57BL6/129Sv genetic background, fewer mice deficient for the tumour-suppressor gene *Trp53* develop testicular tumours than in a pure 129/Sv background⁸. We observed no testicular tumours in 15 *Stra8*-EGFP/*Rosa26* males with a mixed FVB/C57BL6/129Sv genetic background.

The same *Stra8*-EGFP⁺ cells that had been tested for their contribution to spermatogenesis were used to establish the conditions required to convert these cells into pluripotent stem cells. The GFP⁺ SSCs were propagated in basic medium (condition I) for two weeks and then cultured under four different conditions (Supplementary Fig. 1). In condition I, most of the colonies formed

were similar to cultured epiblast cells (Fig. 1a). However, when the cells were cultivated in medium containing leukaemia inhibitory factor (LIF; condition II) or grown on mitomycin-C-treated mouse embryonic fibroblasts (MEFs; condition III), tightly packed embryonic stem cell (ESC)-like colonies appeared (Fig. 1b). These ESC-like colonies can be expanded under standard ESC culture conditions (both MEFs and LIF applied; condition IV). We successfully established four cell lines (SSC5, SSC6, SSC10 and SSC15) from fifteen mice (27% success rate). Two cell lines, SSC5 and SSC6, have now been passaged for more than 30 passages (Fig. 1c).

The phenotype of our cultured GFP⁺ SSCs depends on the culture conditions. Cultured GFP⁺ SSCs expressed the cell-surface marker SSEA-1 (stage-specific embryonic antigen-1; ref. 9) and the germline-specific transcription factor Oct3/4 (ref. 10) (Fig. 1d–f and Supplementary Fig. 2) that characterize undifferentiated mouse ESCs. In common with ESCs, the SSCs in conditions III and IV did not express SSEA-3 (ref. 11), but some cells in conditions I or II stained positively for SSEA-3 (Fig. 1g and Supplementary Fig. 2b). c-Kit is expressed in differentiating spermatogonia but not in SSCs¹², and more cells stained positively for c-kit in conditions I and II than in conditions III and IV (Supplementary Fig. 2b). *Thy1* is a surface marker of SSCs¹³, and can also be detected on ESCs¹⁴. GFP⁺ cells cultured in the presence of LIF (conditions II and IV) are negative for *Thy1*. Like ESCs, they show only low positive staining for the lymphocyte surface marker *Sca1*, and are negative for *Ter119* and *CD34*. Alkaline phosphatase is highly expressed in ESCs but not in SSCs². In our study, alkaline phosphatase was strongly expressed in the GFP⁺ cells in condition IV (Fig. 1h), but a ‘mixed colony’ phenotype was revealed in conditions I–III (Fig. 1i).

These results suggest that SSCs respond to culture conditions^{15,16}, and acquire ESC properties under standard ESC culture conditions. We therefore name these cultured ES-like cells multipotent adult germline stem cells (maGSCs), to distinguish them from SSCs. Cytogenetic analysis showed that maGSCs have a normal karyotype (40 chromosomes, XY) in 80–90% of metaphase spreads. Additionally, and similar to ESCs, maGSCs express higher levels of *Ras1* (also known as *Rasd1*) and *Gro1* (the chemokine *Cxcl1*), but lower levels of *Mcm4* compared to the teratocarcinoma cell line F9 (derived from teratocarcinoma of testis from the 129 mouse strain) and embryonic carcinoma cell (ECC) line P19 (derived from teratocarcinoma of a C3H/He mouse embryo) (Supplementary Fig. 3a), consistent with previous studies¹⁷.

Polymerase chain reaction with reverse transcription (RT–PCR) analysis showed that under all four culture conditions, the cultured cells expressed the transcription factors *Oct3/4* (also known as *Pou5f1*), *Nanog* (ref. 18), *Utf1* (ref. 19), *Esg1* (also known as *Dppa5*; ref. 20) and *Rex1* (also known as *Zfp42*; ref. 21) (Fig. 1j), similar to

¹Department of Cardiology and Pneumology, Heart Center, Georg-August-University of Göttingen, Robert-Koch-Str. 40, 37075 Göttingen, Germany. ²Institute of Human Genetics and ³Department of Cellular and Molecular Immunology, Georg-August-University of Göttingen, Heinrich-Dücker-Weg 12, 37073 Göttingen, Germany.

*These authors contributed equally to this work.

ESCs. Notably, cells cultured with LIF (conditions II and IV) expressed higher levels of *Stra8* than those cultured without LIF (conditions I and III).

On the basis of our experience with isolating SSCs from the GFP⁺ transgenic mice, we were able to identify the cells morphologically, which allowed us to rule out the possibility that the SSC isolation was restricted to the transgenic mice. We have successfully derived ES-like cell lines from mice with three different genotypes (C57BL/6, 129/Ola and FVB) without genetic selection. ES-like cells were isolated from

two out of eleven C57BL/6 mice (18%), one out of seven FVB mice (14%) and one out of three 129/Ola (33%) mice. These cultures, similar to the GFP⁺ cells, express *Stra8*, *Oct3/4*, *Nanog*, *Utf1*, *Esg1* and *Rex1*, and are susceptible to LIF (Supplementary Fig. 3b).

Analysis of the DNA microsatellite profile of the established cultures rules out the possibility of contamination with ESCs, ECCs or MEFs cultured in the same facility (Supplementary Tables 1–4). Furthermore, we show that the two lines from the C57BL/6 inbred strain differ in the four DNA microsatellite markers, indicating they are two separate lines (Supplementary Table 1).

To induce *in vitro* differentiation of maGSCs, we applied the ‘hanging drop’ method used for ESC differentiation²². We observed an overall decrease in GFP⁺ cell populations upon embryoid body differentiation (Supplementary Fig. 4). In parallel, the expression of *Oct3/4*, *Nanog*, *Utf1*, *Esg1* and *Rex1* gradually decreased during differentiation (Fig. 1k). To determine whether maGSCs can spontaneously differentiate into derivatives of the three primary germ layers, we examined the expression of a panel of cell-specific genes and proteins during embryoid body differentiation (Figs 2–4).

Differentiation of mesodermal lineages (for example, cardiac, skeletal muscle and vascular cells) was confirmed by expression of the early mesoderm marker *brachyury* (*T*)²³ as well as lineage-specific genes and proteins. The expression of *brachyury* is maximal at early differentiation stages (Fig. 2). The cardiac transcription factors *islet-1* (ref. 24), *Gata4*, *Nkx2.5* and *Mef2c* are already expressed at high levels at day 5, whereas cardiac-specific genes such as α -myosin heavy chain (α -MHC, known as *Myh6*), ventricular isoform 2 of myosin light chain (*Mlc2v*, also known as *Myl2*) and atrial natriuretic factor (*ANF*, also known as *Nppa*) are expressed at high levels two days later

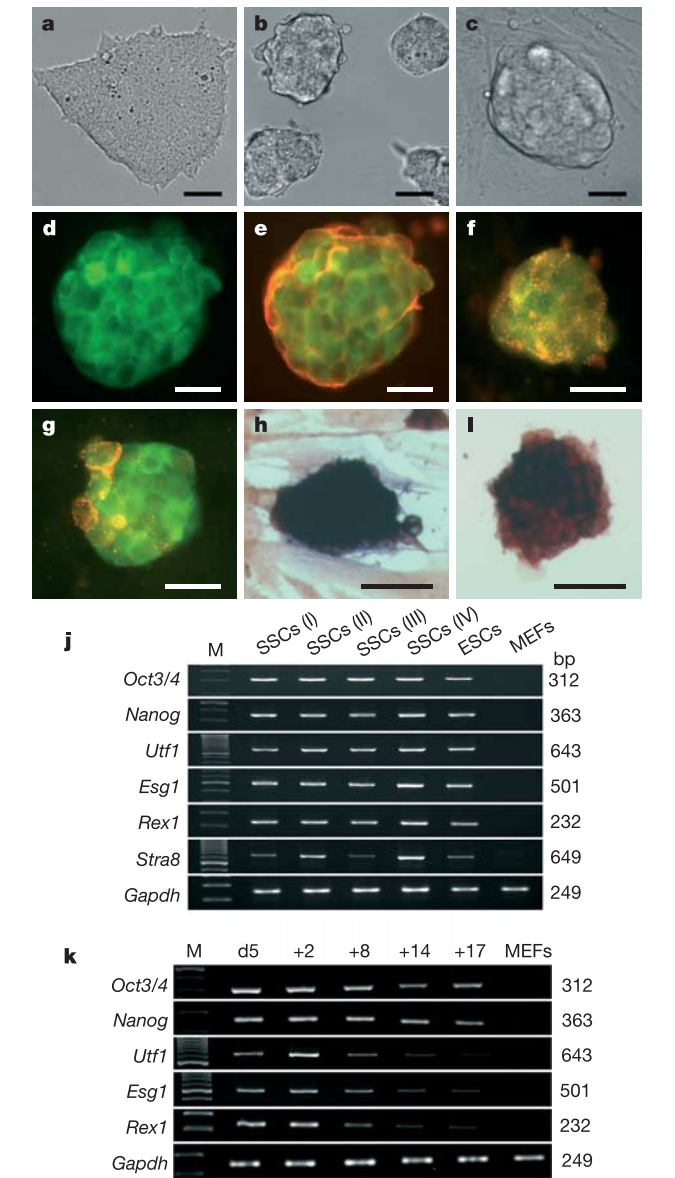


Figure 1 | Cellular and molecular characterization of cultured SSCs and maGSCs. **a**, Epiblast-like colony formed under culture condition I. **b**, ESC-like colonies appeared under culture condition II. **c**, A typical colony of established culture under condition IV at passage 30. **d–g**, Double immunostaining of maGSCs in culture condition IV (**d–f**) or condition II (**g**) with antibodies against GFP (green, **d–g**) and SSEA-1 (red, **e**), Oct4 (red, **f**) or SSEA-3 (red, **g**). **h**, **i**, Alkaline phosphatase staining. SSCs cultured under condition IV (maGSCs, **h**) are strongly positive for alkaline phosphatase, whereas SSCs under condition II (**i**) show a mixed colony phenotype. **j**, **k**, RT-PCR analyses of transcription factors essential for undifferentiated cells in SSCs cultured under conditions I, II, III and IV (**j**) and during differentiation of embryoid bodies after plating at day 5 (d5; **k**). M, 100-bp DNA markers; numbers to the right indicate the sizes of the resolved DNA fragments (in bp). Scale bars, 50 μ m (**a–c**, **h**, **i**), 25 μ m (**d–g**).

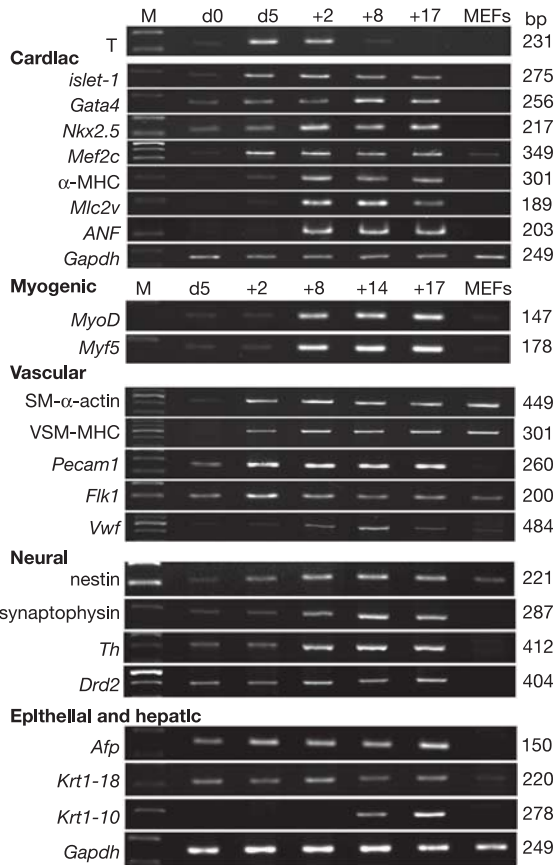


Figure 2 | RT-PCR analysis of lineage-specific transcription factors and genes. Analyses were performed at different stages during the differentiation of embryoid bodies after plating at day 5 (d5). M, 100-bp DNA markers. d0, maGSCs before embryoid body formation.

(Fig. 2). Spontaneously contracting cells appeared as clusters (Supplementary Video 1) and were identified in approximately 90% of the embryoid bodies ($n = 144$; three independent experiments) at day 5 + 5 (at day 5 after plating of 5-day-old embryoid bodies), comparable to the cardiac differentiation of ESCs (86%). The cardiomyocyte phenotype of these contracting areas can be shown by immunostaining for proteins relevant for myocyte contraction. Single cardiomyocytes isolated from beating areas show sarcomeric striations when stained for α -sarcomeric actinin (Fig. 3a), sarcomeric MHC and cardiac troponin T, organized in bundles (Fig. 3b, c). Expression of the gap-junction protein connexin 43 at cell-to-cell contacts in cardiac clusters (Fig. 3d–f) indicates cell-to-cell communication. In addition, single cardiomyocytes show spontaneous action potentials, as analysed using patch-clamp electrophysiology (Fig. 3g).

The differential expression of cardiac- and skeletal-muscle-specific transcription factors makes it possible to distinguish maGSC-derived cardiomyocytes from skeletal muscle cells. The myogenic regulatory

factors *MyoD* and *Myf5* are strongly expressed from day 5 + 8 (Fig. 2), whereas cardiac-specific transcription factors are strongly expressed by day 5. Moreover, the first myoblasts positive for nebulin²⁵ appear at day 5 + 7, whereas cardiac cells are never positive for nebulin. The formation of multinucleated contracting myotubes positive for nebulin was only observed at late stages (Fig. 3h). These data show that the cardiac and skeletal muscle lineages are different from early on in the differentiation process.

The differentiation of vascular endothelial and smooth muscle cells was confirmed by the expression of genes encoding smooth muscle α -actin (SM- α -actin, known as *Acta2*), vascular smooth muscle MHC (VSM-MHC, known as *Myh11*), platelet/endothelial-cell adhesion molecule 1 (*Pecam1*), *Flk1* (also known as *Kdr*) and von Willebrand factor (*Vwf*) (Fig. 2). In embryoid body outgrowths we found cells positive for both the uptake of DiI-conjugated acetylated low-density lipoproteins (ac-LDL-DiI; (Fig. 3i) and lectin binding (Fig. 3j), which are characteristic of endothelial cells²⁶. We also observed the formation of cellular networks and tube-like structures in embryoid body outgrowths at later stages using antibodies against Vwf (Fig. 3k) and smooth muscle α -actin (Fig. 3l), indicating that these net-like and tubular structures consist of vascular endothelial and smooth muscle cells.

Neuroectoderm differentiation was confirmed by the expression of

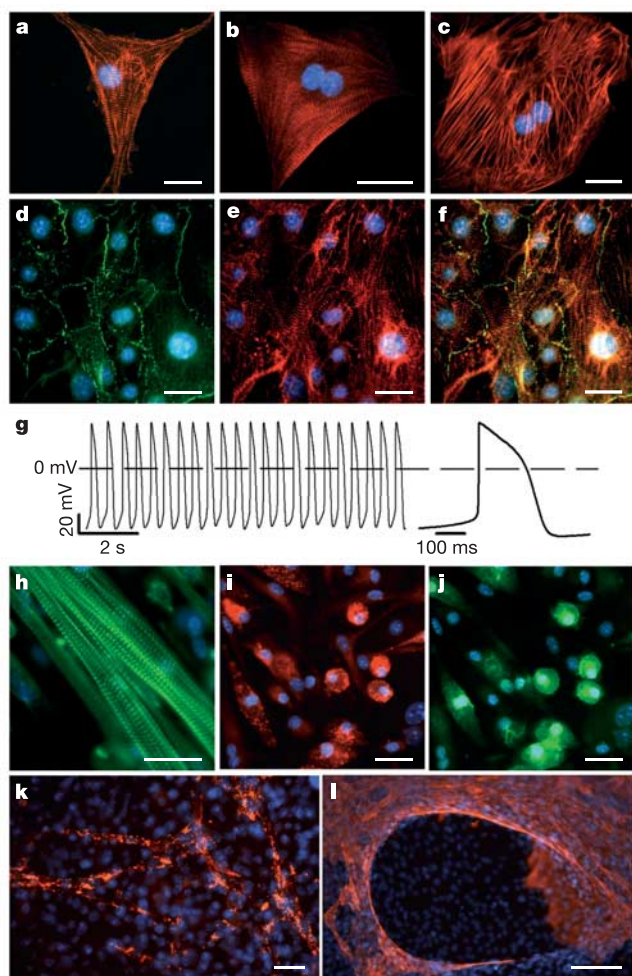


Figure 3 | Mesoderm differentiation of maGSCs. **a–c**, Organization of the sarcomeric proteins α -actinin (**a**), sarcomeric MHC (**b**) and cardiac troponin T (**c**) in isolated cardiomyocytes at day 5 + 7. **d–f**, Connexin 43 staining (**d**, green) in a cluster of uninucleate cardiac cells stained for sarcomeric α -actinin (**e**, red), with an overlay of **d** and **e** shown in **f**. Nuclei are stained with DAPI. **g**, Original traces of ventricle-like action potentials in a cardiomyocyte derived from maGSCs. **h**, Nebulin-positive myotubes at differentiation day 5 + 23. **i**, **j**, DiI-acLDL uptake (**i**, red) and lectin binding (**j**, green) of endothelial cells in embryoid body outgrowths at day 5 + 14. **k**, Vwf-positive endothelial cells (red) at day 5 + 17. **l**, Smooth muscle α -actin-positive cells (red) of tube-like structure in embryoid body outgrowths at day 5 + 14. Scale bars, 25 μ m (**a–f**, **h–k**), 100 μ m (**l**).

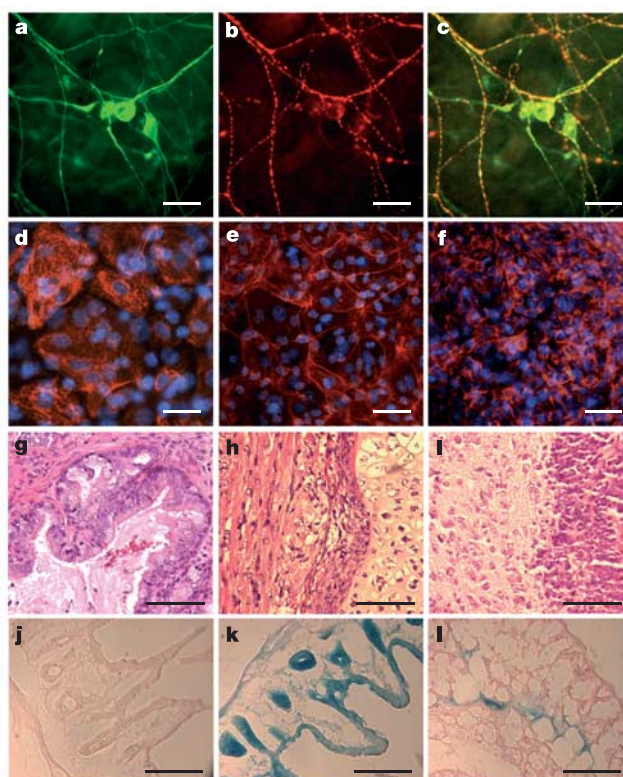


Figure 4 | Differentiation of maGSCs *in vitro* and *in vivo*. **a–c**, Neuronal differentiation of maGSCs. TH-positive dopaminergic neurons (**b**) appeared among neurofilament protein M (NFM)-positive neurons (**a**) on day 12 after plating of embryoid bodies at day 5 (day 5 + 12). **c**, Overlay of **a** and **b**. **d–f**, Epithelial cell/hepatocyte differentiation. **d**, Pan-cytokeratin-positive epithelial cells on day 5 + 14. **e**, CK18-positive large epithelioid cells on day 5 + 17. **f**, CK7-positive bile duct cells on day 5 + 14. Nuclear staining with DAPI. **g–i**, Teratomas from maGSCs. The tumours contained abundant differentiation of advanced derivatives of all three embryonic germ layers. Shown are epithelium with intestinal differentiation (**g**), striated muscle, cartilage (**h**) and neural tissue (**i**). **j–l**, Representative images of LacZ-stained intestine tissue from a wild-type mouse (**j**), a mouse with high grade of chimaerism (**k**) and a mouse with low grade of chimaerisms (**l**). Scale bars, 25 μ m (**a–f**), 50 μ m (**g–i**), 100 μ m (**j–l**).

nestin, a marker for neuroepithelial precursors, and of synaptophysin, tyrosine hydroxylase (*Th*) and dopamine receptor 2 (*Drd2*), markers for differentiated neurons (Fig. 2). We found high numbers of nestin-positive neuroprogenitors at day 5 + 5 (data not shown). Although the spontaneous differentiation of maGSCs into neuronal cells (Fig. 4a–c) was low (only 25% of embryoid bodies contained neuronal cells at day 5 + 11), maGSCs could differentiate into dopaminergic neurons under standard embryoid body differentiation conditions (Fig. 4a–c).

We characterized epithelial-like cells in embryoid bodies using an anti-pan cytokeratin antibody, which reacts with a wide variety of epithelial tissues (Fig. 4d). Cytokeratin 10 (*Krt1-10*), which is present in keratinizing stratified epithelia, was expressed at late differentiation stages (Fig. 2). We also found cells with phenotypic characteristics of hepatocytes in embryoid bodies. Expression of early (α fetoprotein, *Afp*) and late (*Krt1-18*) markers of hepatocyte differentiation was observed during embryoid body differentiation (Fig. 2). At day 5 + 8, cells expressing AFP have one of two different morphologies: round or spindle shapes. At day 5 + 17, many large epithelioid cells stained positive for cytokeratin 18 (CK18) (Fig. 4e). In addition, we found many cells positive for CK7, a maturation marker in bile duct cells (Fig. 4f).

To confirm the pluripotency of maGSCs *in vivo*, we subcutaneously injected the cells into SCID/beige mice. The transplanted cells formed mature teratomas in all recipients (10/10) by 6 weeks after inoculation. The teratomas contained derivatives of three embryonic germ layers, including epithelium with intestinal differentiation (endoderm), striated muscle, smooth muscle, fat, bone and cartilage (mesoderm), and neural tissue (ectoderm) (Fig. 4g–i).

To determine the developmental potential of SSCs, we injected 10–15 GFP⁺ SSCs into 3.5-day-old blastocysts. Sixty-five injected blastocysts were transferred into the uterus of pseudopregnant mice. The number of litters born (42 animals) and number of animals per litter (6–7) were consistent with the birth rate seen with ESC implantation²⁷. Animals born from microinjected blastocysts were of similar size as normal animals and did not have overt abnormalities. Chimaerism could be detected in 39 of 42 of mice (~93%) by *LacZ* PCR of genomic DNA isolated from tail biopsies (Supplementary Fig. 5a). We identified a contribution of *LacZ*⁺ SSCs to many tissues, including heart, brain, intestine, lung, skeletal muscle, liver, kidney, spleen and testis, by both PCR analysis (Supplementary Fig. 5b) and *LacZ* staining (Fig. 4j–l, Supplementary Fig. 6). In addition, six male chimaeric mice were each mated with two female chimaeric mice for 2–3 weeks. Two litters (from two different males) were born. DNA analysis from one litter (6 animals) showed three animals positive for *LacZ* and two animals positive for GFP (segregation of the locus in which *Gfp* and *LacZ* integrated; Supplementary Fig. 5c), indicating germline transmission.

The most striking observation from our experiments is the pluripotency of adult SSCs and SSC-derived maGSCs, similar to ESCs. A previous study failed to establish ES-like cells from adult mouse testis². Here we used a genetic selection⁵ to enrich SSCs, which is more efficient than CD9 antibody selection^{2,28}. Unlike the previous study, we did not apply basic fibroblast growth factor (bFGF), epidermal growth factor (EGF) and LIF for the initial culture of testicular cells, growth factors that may result in the overgrowth of other cells, including fibroblasts, endothelial and Sertoli cells. Although GDNF is essential for the self-renewal of SSCs *in vivo*²⁹, we found that the proliferation of mouse SSCs *in vitro* is not dependent on GDNF but instead on LIF. Generation of ESC-like maGSCs *in vitro* may result from genetic reprogramming of SSCs in culture, similar to that which occurs during the 'cloning process'³⁰. Alternatively, SSCs themselves may be multipotent. In this case, under *in vivo* physiological conditions, the interaction with Sertoli cells in the testis may lead SSCs to spermatogenesis and inhibit multi-lineage differentiation. However, when SSCs are expanded in the absence of Sertoli cells, as in our *in vitro* culture and in blastocysts,

some SSCs might be released from this inhibition and converted into pluripotent stem cells. This is similar to the conversion of embryonic germ cells from primordial germ cells¹. Cell fusion has been suggested as an explanation for stem-cell plasticity. Our *in vitro* studies show the multi-lineage differentiation potential of maGSCs in the absence of cell fusion, as no co-culture system is applied.

Human SSCs may have great potential for cell-based organ regeneration therapy and for studying genetic diseases in various cell lineages. Therefore, development of a culture system for establishing human maGSCs from testicular biopsies is of paramount importance.

METHODS

Isolation and culture of SSCs. SSCs were isolated from testis tissues of adult transgenic (Stra8-EGFP/Rosa26) and inbred (C57BL/6, FVB and 129/Ola) mice (4–6 weeks old). Details are available in the Supplementary Information.

GFP⁺ cells from Stra8-EGFP/Rosa26 mice were separated by FACS. To maintain the GFP⁺ cells in an undifferentiated state, four different culture conditions were tested. For condition I, cells were cultured on gelatine-coated culture dishes in basic medium (DMEM medium supplemented with 15% FCS, 2 mM L-glutamine (Gibco), 50 μ M β -mercaptoethanol (Promega), 1 $\times$ nonessential amino acids (NEAA, Gibco)). For condition II, cells were cultured on gelatine-coated culture dishes in basic medium containing 10³ units ml⁻¹ LIF (ESGRO, Chemicon). For condition III, cells were cultured on MEFs in basic medium. For condition IV, cells were cultured on MEFs with basic medium containing LIF (ESC culture condition).

Details of SSC transplantation into seminiferous tubules of testes, cytogenetic analysis, *in vitro* differentiation using the 'hanging drop' method, teratoma formation, chimaeras and germline transmission, isolation of single cardiomyocytes, immunocytochemistry, alkaline phosphatase staining, RT-PCR, action potential recordings and *LacZ* staining are presented in the Supplementary Methods.

Received 21 November 2005; accepted 7 March 2006.

Published online 24 March 2006.

1. Matsui, Y., Zsebo, K. & Hogan, B. L. Derivation of pluripotent embryonic stem cells from murine primordial germ cells in culture. *Cell* **70**, 841–847 (1992).
2. Kanatsu-Shinohara, M. *et al.* Generation of pluripotent stem cells from neonatal mouse testis. *Cell* **119**, 1001–1012 (2004).
3. Spradling, A., Drummond-Barbosa, D. & Kai, T. Stem cells find their niche. *Nature* **414**, 98–104 (2001).
4. Oulad-Abdelghani, M. *et al.* Characterization of a premeiotic germ cell-specific cytoplasmic protein encoded by *Stra8*, a novel retinoic acid-responsive gene. *J. Cell Biol.* **135**, 469–477 (1996).
5. Giull, G. *et al.* Murine spermatogonial stem cells: targeted transgene expression and purification in an active state. *EMBO Rep.* **3**, 753–759 (2002).
6. Nayernia, K. *et al.* Stem cell based therapeutical approach of male infertility by teratocarcinoma derived germ cells. *Hum. Mol. Genet.* **13**, 1451–1460 (2004).
7. Stevens, L. C. & Hummel, K. P. A description of spontaneous congenital testicular teratomas in strain 129 mice. *J. Natl. Cancer Inst.* **18**, 719–747 (1957).
8. Donehower, L. A. *et al.* Effects of genetic background on tumorigenesis in p53-deficient mice. *Mol. Carcinog.* **14**, 16–22 (1995).
9. Solter, D. & Knowles, B. B. Monoclonal antibody defining a stage-specific mouse embryonic antigen (SSEA-1). *Proc. Natl. Acad. Sci. USA* **75**, 5565–5569 (1978).
10. Nichols, J. *et al.* Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95**, 379–391 (1998).
11. Damjanov, I. *et al.* Immunohistochemical localization of murine stage-specific embryonic antigens in human testicular germ cell tumors. *Am. J. Pathol.* **108**, 225–230 (1982).
12. Rossi, P., Sette, C., Dolci, S. & Geremia, R. Role of c-kit in mammalian spermatogenesis. *J. Endocrinol. Invest.* **23**, 609–615 (2000).
13. Kubota, H., Avarbock, M. R. & Brinster, R. L. Culture conditions and single growth factors affect fate determination of mouse spermatogonial stem cells. *Biol. Reprod.* **71**, 722–731 (2004).
14. Ling, V. & Neben, S. *In vitro* differentiation of embryonic stem cells: immunophenotypic analysis of cultured embryoid bodies. *J. Cell. Physiol.* **171**, 104–115 (1997).
15. Nagano, M., Ryu, B. Y., Brinster, C. J., Avarbock, M. R. & Brinster, R. L. Maintenance of mouse male germ line stem cells *in vitro*. *Biol. Reprod.* **68**, 2207–2214 (2003).
16. Kubota, H., Avarbock, M. R. & Brinster, R. L. Growth factors essential for self-renewal and expansion of mouse spermatogonial stem cells. *Proc. Natl. Acad. Sci. USA* **101**, 16489–16494 (2004).

17. Bonner, A. E., Wang, Y. & You, M. Gene expression profiling of mouse teratocarcinomas uncovers epigenetic changes associated with the transformation of mouse embryonic stem cells. *Neoplasia* **6**, 490–502 (2004).
18. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631–642 (2003).
19. Okuda, A. *et al.* UTF1, a novel transcriptional coactivator expressed in pluripotent embryonic stem cells and extra-embryonic cells. *EMBO J.* **17**, 2019–2032 (1998).
20. Tanaka, T. S. *et al.* Gene expression profiling of embryo-derived stem cells reveals candidate genes associated with pluripotency and lineage specificity. *Genome Res.* **12**, 1921–1928 (2002).
21. Ben-Shushan, E., Thompson, J. R., Gudas, L. J. & Bergman, Y. *Rex-1*, a gene encoding a transcription factor expressed in the early embryo, is regulated via Oct-3/4 and Oct-6 binding to an octamer site and a novel protein, Rox-1, binding to an adjacent site. *Mol. Cell. Biol.* **18**, 1866–1878 (1998).
22. Guan, K., Rohwedel, J. & Wobus, A. M. Embryonic stem cell differentiation models: cardiogenesis, myogenesis, neurogenesis, epithelial and vascular smooth muscle cell differentiation *in vitro*. *Cytotechnology* **30**, 211–226 (1999).
23. King, T., Beddington, R. S. & Brown, N. A. The role of the brachyury gene in heart development and left-right specification in the mouse. *Mech. Dev.* **79**, 29–37 (1998).
24. Laugwitz, K. L. *et al.* Postnatal *Isl1*⁺ cardioblasts enter fully differentiated cardiomyocyte lineages. *Nature* **433**, 647–653 (2005).
25. Furst, D. O., Osborn, M. & Weber, K. Myogenesis in the mouse embryo: differential onset of expression of myogenic proteins and the involvement of titin in myofibril assembly. *J. Cell Biol.* **109**, 517–527 (1989).
26. Dimmeler, S. *et al.* HMG-CoA reductase inhibitors (statins) increase endothelial progenitor cells via the PI 3-kinase/Akt pathway. *J. Clin. Invest.* **108**, 391–397 (2001).
27. Schoonjans, L. *et al.* Improved generation of germline-competent embryonic stem cell lines from inbred mouse strains. *Stem Cells* **21**, 90–97 (2003).
28. Kanatsu-Shinohara, M., Toyokuni, S. & Shinohara, T. CD9 is a surface marker on mouse and rat male germline stem cells. *Biol. Reprod.* **70**, 70–75 (2004).
29. Meng, X. *et al.* Regulation of cell fate decision of undifferentiated spermatogonia by GDNF. *Science* **287**, 1489–1493 (2000).
30. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. Viable offspring derived from fetal and adult mammalian cells. *Nature* **385**, 810–813 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. A summary figure is also included.

Acknowledgements We thank A. Cierpka, D. Meyer, S. Wolf, B. Sadowski, I. Schwandt, C. Müller and S. Burkhardt for technical assistance. We thank M. Schindler, H. Riedesel and S. Wolf for their assistance in the generation of transgenic mice, G. Wulf for help with FACS analysis, and B. Hemmerlein for the analysis of teratomas. This work was supported by grants from the Georg-August-University of Göttingen (Forschungsförderungsprogramm Stammzellen) to K.G. and K.N., and a DFG grant (Emmy-Noether Program) to L.S.M.

Author Contributions K.G., G.H., K.N. and W.E. conceived and designed the experiments and performed the data analysis and controls. K.G., K.N., L.S.M., S.W., R.D., J.H.L., J.N., F.W. and M.L. performed the experiments, and K.G., G.H., K.N. and W.E. wrote the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.H. (hasenfus@med.uni-goettingen.de) or W.E. (wengel@gwdg.de).

LETTERS

Recursive syntactic pattern learning by songbirds

Timothy Q. Gentner^{1†}, Kimberly M. Fenn², Daniel Margoliash^{1,2} & Howard C. Nusbaum²

Humans regularly produce new utterances that are understood by other members of the same language community¹. Linguistic theories account for this ability through the use of syntactic rules (or generative grammars) that describe the acceptable structure of utterances². The recursive, hierarchical embedding of language units (for example, words or phrases within shorter sentences) that is part of the ability to construct new utterances minimally requires a 'context-free' grammar^{2,3} that is more complex than the 'finite-state' grammars thought sufficient to specify the structure of all non-human communication signals. Recent hypotheses make the central claim that the capacity for syntactic recursion forms the computational core of a uniquely human language faculty^{4,5}. Here we show that European starlings (*Sturnus vulgaris*) accurately recognize acoustic patterns defined by a recursive, self-embedding, context-free grammar. They are also able to classify new patterns defined by the grammar and reliably exclude agrammatical patterns. Thus, the capacity to classify sequences from recursive, centre-embedded grammars is not uniquely human. This finding opens a new range of complex syntactic processing mechanisms to physiological investigation.

The computational complexity of generative grammars is formally defined³ such that certain classes of temporally patterned strings can only be produced (or recognized) by specific classes of grammars (Fig. 1). Starlings sing long songs composed of iterated motifs (smaller acoustic units)⁶ that form the basic perceptual units of individual song recognition^{7–9}. Here we used eight 'rattle' and eight 'warble' motifs (see Methods) to create complete 'languages' (4,096 sequences) for two distinct grammars: a context-free grammar (CFG) of the form A^2B^2 that entails recursive centre-embedding, and a finite-state grammar (FSG) of the form $(AB)^2$ that does not (Fig. 2a, b; 'A' refers to rattles and 'B' to warbles).

We trained 11 European starlings, using a go/nogo operant conditioning procedure, to classify subsets of sequences from these languages (see Methods and Supplementary Information). Nine out of eleven starlings learned to classify the FSG and CFG sequences accurately (as assessed by d' , which provides an unbiased measure of sensitivity to differentiating between two classes of patterns), but this task was difficult (Fig. 2c). The rate of acquisition varied widely among the starlings that learned the task (303.44 ± 57.11 blocks to reach criterion (mean $\pm$ s.e.m.), range 94–562 blocks with 100 trials per block), and was slow by comparison to other operant song-recognition tasks⁷.

To assess the possibility that starlings learned to classify correctly the motif patterns described by the CFG and FSG grammars through rote memorization of the training exemplars, we further tested the first four birds to reach stable asymptotic performance on the initial classification training (mean $d' \pm$ s.e.m. at asymptote 2.52 ± 0.40 , Fig. 2d). We transferred the birds abruptly from the 16 baseline training stimuli to 16 new sequences from the same two grammars (A^2B^2 and $(AB)^2$, eight sequences from each) while maintaining the same operant contingencies used during baseline training. Starlings

correctly classified the new CFG and FSG sequences during the first transfer session (Fig. 3a). The mean d' over the first 100 trials with new stimuli (roughly six responses to each exemplar) was 1.08 ± 0.50 , which is significantly better than chance performance ($d' = 0$). Over the first five 100-trial blocks of the transfer session, the mean d' was 1.14 ± 0.20 (Fig. 3a), and the lower bound of the 95% confidence interval (CI) around d' was above zero for all birds (range 0.34–1.18), with subsequent performance continuing to improve. Thus, the birds did not simply memorize the 16 baseline training stimuli, but instead acquired general knowledge about features diagnostic of the two grammars, and applied this knowledge to classify the new stimuli correctly. Given that the same elements (motifs) composed the sequences in each class, this knowledge must be related to the differential patterning of these elements by each grammar. Additional generalization tests using 'probe' procedures that test for learning during exposure to the new grammatical stimuli (see Methods and Fig. 3b) also reject the rote memorization hypothesis, and support the conclusion that the birds acquired information about the patterning of motifs in the CFG and FSG classes.

One possibility consistent with interpretations of experiments on syntactic processing in cotton-top tamarins^{10–12} is that the birds learned only the FSG, and treated the grammatical CFG sequences as the complement set. However, the results of further probe tests rule out this possibility. We constructed 16 motif sequences based on four different agrammatical patterns (AAAA, BBBB, ABBA and BAAB, with four exemplars per pattern, using the same A and B motifs as for the FSG and CFG grammars) and presented them along with new A^2B^2 and $(AB)^2$ patterns as probe stimuli (Methods). The response

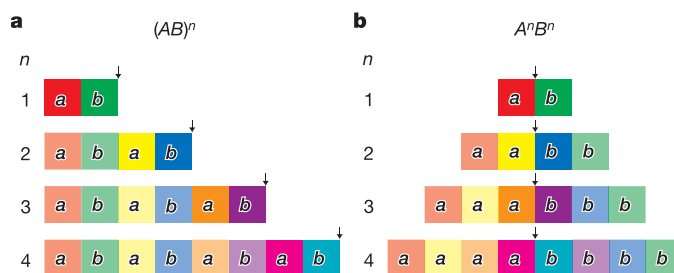


Figure 1 | Grammatical forms. **a**, Finite-state form $(AB)^n$. **b**, Context-free form A^nB^n . Both grammars describe patterned sequences of elements (lower-case letters) of the sets 'A' and 'B'. Longer strings of the form $(AB)^n$, where n gives the number of AB iterations, are produced by appending elements to the end of an $n - 1$ sequence. Longer strings with the form A^nB^n are produced by embedding elements into the centre of an $n - 1$ sequence. Learning of and generalization to an A^nB^n pattern implies the capacity to process syntactic structures generated through recursive centre-embedding. Black arrows denote insertion points for higher-order sequences. Brightly coloured squares mark the 'AB' phrase inserted at each order. Different hues denote different elements.

¹Department of Organismal Biology and Anatomy and ²Department of Psychology, University of Chicago, Chicago, Illinois 60637, USA. [†]Present address: Department of Psychology, University of California, San Diego, La Jolla, California 92093, USA.

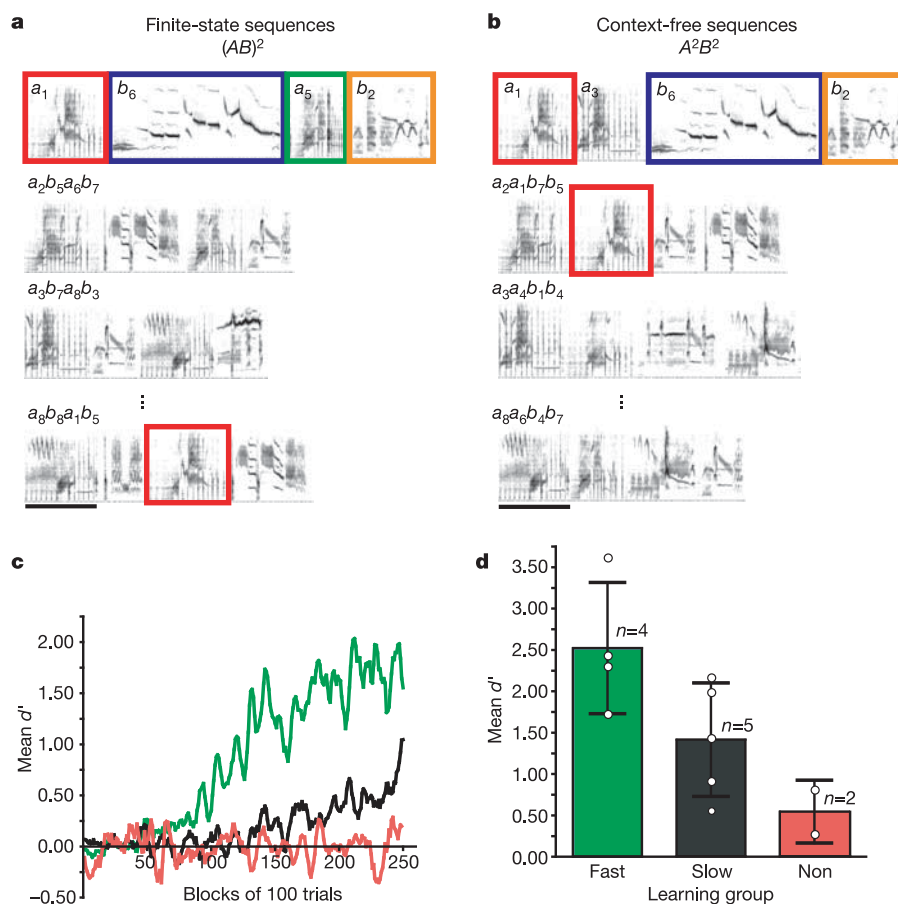


Figure 2 | Classification of grammatical pattern stimuli. **a, b**, Sonograms (frequency range 0.2–10.0 kHz; scale bars, 1 s) showing four of the eight sequences constructed from the finite-state grammar $(AB)^n$ (**a**) and the context-free grammar A^nB^n (**b**) used in the initial FSG versus CFG pattern classification training with $n = 2$. Similarly coloured boxes mark the same motifs in multiple sequences. The position of a motif within a sequence is arbitrary with respect to its subscript label. See Supplementary Information for complete stimulus patterns and sonograms. **c**, Acquisition

patterns for the agrammatical probe stimuli differed significantly from those for new $(AB)^2$ stimuli for all four birds, and from those for new A^2B^2 stimuli for three of the four birds (X^2 calculated separately for each bird and stimulus class, $P < 0.05$ in 7 of 8 cases; $X^2 = 1.74$ (not significant) for one case). It may be relevant that the nonconforming bird was the only one of the four tested for which the FSG sequences served as the S^+ stimuli (see Methods). Regardless, three of the four birds clearly learned to classify both the CFG and FSG stimuli during training, suggesting that they learned both FSG and the CFG patterning rules.

Time and memory capacity both constrain the functional length of any grammatical string, yet part of the power of a generative grammar is its capacity to describe strings of arbitrary length. To test whether our birds generalized from A^2B^2 and $(AB)^2$ to higher orders of grammatical structure, we probed the birds with $n = 3$ (that is, A^3B^3 and $(AB)^3$) and $n = 4$ motif sequences while they maintained baseline ($n = 2$) classification (see Methods and Fig. 3b). All birds accurately classified the $n = 3$ CFG and FSG sequences (mean $d' = 1.37 \pm 0.54$; range for lower bound of 95% CI 0.03–2.23) and the $n = 4$ CFG and FSG sequences (mean $d' = 1.27 \pm 0.22$; range for lower bound of 95% CI 0.23–1.54). Thus, classification training with $n = 2$ sequences can induce behaviour consistent with higher-order, generative grammars, including those using recursive centre-embedding.

An alternative explanation for these results is that the birds learned

curves for the baseline FSG/CFG classification, showing mean d' over the first 250 blocks (100 trials per block) for birds that learned quickly and were subjected to further testing (green), birds that learned slowly (black) and birds that did not reach the accuracy criterion (red; see Methods). **d**, Mean d' ($\pm$ s.d.) on the baseline CFG versus FSG classification task at asymptote. Open circles show means from individual birds. Colours and groups as in c.

a 'simpler' grammar that approximates the recursive structure in the A^nB^n sequences. Sequences that follow the pattern A^nB^n constitute a subset of those that follow the more general pattern A^*B^* , in which the number of a 's and b 's can vary independently. Although a CFG is required to produce sequences in which the number of a 's and b 's are matched, (as in A^nB^n), the whole of A^*B^* can be generated by a finite-state grammar. We tested whether the birds learned an A^*B^* finite-state approximation to A^nB^n by examining their responses to the following A^*B^* patterns: A^1B^3 , A^3B^1 , A^2B^3 and A^3B^2 (four randomly chosen sequences per pattern, same A/B motif vocabularies as all FSG and CFG stimuli). We presented the A^*B^* stimuli along with A^nB^n and $(AB)^n$ ($n = 2, 3, 4$) sequences as probes during the same sessions.

If birds learned A^*B^* as a finite-state approximation to A^nB^n , then the pattern of response to each A^*B^* stimulus should match the response to A^nB^n reference stimuli. Our results reject this hypothesis. All birds showed a strong bias to treat the A^*B^* patterns differently from the A^nB^n reference stimuli, while maintaining accurate classification of the A^nB^n and $(AB)^n$ reference and training stimuli (mean $d' \geq 1.19$ in all four cases). Responses to all four A^*B^* patterns were significantly different from responses to the A^2B^2 and A^4B^4 reference stimuli ($X^2 < 0.001$ for all 8 cases, d.f. = 3). Responses to the A^1B^3 , A^3B^2 and A^3B^1 patterns were significantly different from responses to the A^3B^3 patterns ($X^2 < 0.001$ in all cases, d.f. = 3). In addition, all six pair-wise comparisons between responses to individual A^*B^*

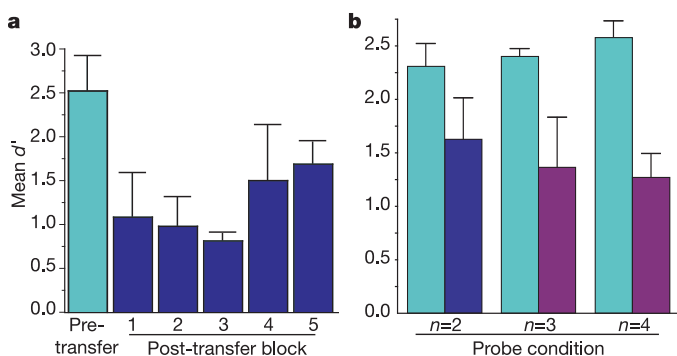


Figure 3 | Generalization to new FSG and CFG sequences. **a**, Mean d' ($\pm$ s.e.m.) for transfer from the training to new FSG and CFG stimuli (turquoise, mean performance over the five blocks of trials preceding transfer; blue, performance in the first five blocks after transfer; 100 trials per block). Performance was stable across these post-transfer blocks ($F_{3,4} = 1.15$, $P = 0.35$, repeated measures ANOVA), then increased gradually to pre-transfer levels (not shown). All mean d' values shown are significantly greater than zero (see text). Acquisition of the transfer stimuli was much faster than for the original training sets (12.50 ± 3.11 blocks to criterion (mean $\pm$ s.e.m.), range 8–15 blocks; 100 trials per block), which can be attributed partially to generalization across the CFG and FSG classes. **b**, Mean d' ($\pm$ s.e.m.) during grammatical probe sessions. Birds correctly classified new A^nB^n and $(AB)^n$ sequences when $n = 2$ (blue), $n = 3$ or $n = 4$ (purple). Classification accuracy was significantly above chance for all three types of probe sequences (mean d' for $n = 2$, 1.63 ± 0.39 ; see text for $n = 3$, $n = 4$). Classification of the baseline training stimuli (turquoise) was well above chance for all three conditions (mean $d' \geq 2.39$, s.d. ≤ 0.25). The drop between training and probe stimulus classification was significant in only the $n = 4$ condition ($P < 0.05$, Mann–Whitney U -test), suggesting that these sequences were more difficult to classify correctly than the other grammatical test sequences (see Supplementary Information).

patterns were significantly different ($\chi^2 < 0.0005$ for all cases, d.f. = 3) suggesting that the birds did not treat the A^nB^n patterns as a single stimulus class. These results suggest that the birds did not solve the recursive classification tasks by learning a finite-state approximation to the CFG. Rather, it seems that they learned A^nB^n , or a functionally equivalent rule.

We then used the pattern of responses to the various agrammatical probe stimuli to test alternative hypotheses for the starlings' accurate classification of A^nB^n and $(AB)^n$ patterns. For example, the task could reduce to the classification 'AA' and 'AB', or 'BB' and 'AB', if only the initial (primacy) or terminal (recency) motif pairs are attended to, respectively. Similarly, birds could correctly classify A^nB^n and $(AB)^n$ motif sequences by listening for B/A transitions (A^nB^n patterns have none), counting the number of A/B transitions (A^nB^n patterns have only one), or listening for AA (or BB) motif pairs. Each of these potential solution strategies can be and were ruled out by considering specific comparisons among the various agrammatical probe stimuli (Fig. 4 and Supplementary Information). In all cases, classification of the agrammatical patterns was significantly worse (lower d') than grammatical probe stimuli, suggesting that these alternative strategies were not the basis for generalization performance. Instead, starlings seem to have learned the patterning rules defined by each grammar.

We thus demonstrate that starlings can recognize syntactically well-formed strings, including those that use a recursive centre-embedding rule. At least a simple level of recursive syntactic pattern processing is therefore shared with other animals. These results challenge the recent claim that recursion forms the computational core of a uniquely human narrow faculty for language (FLN)⁴. We attempted to rule out the most plausible finite-state solution strategies that could account for accurate classification of A^nB^n patterns (Supplementary Information), suggesting that the learned patterning rule conforms to a stochastic context-free grammar. In

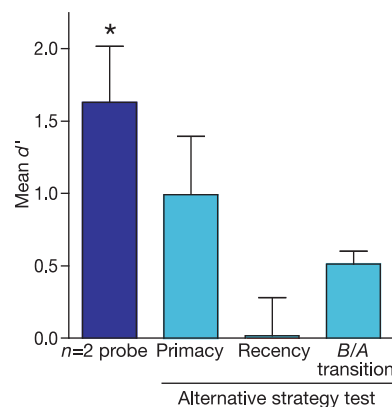


Figure 4 | Agrammatical controls for alternative strategies. Mean d' ($\pm$ s.e.m.) values for comparisons among the AAAA, BBBB, ABBA and BAAB agrammatical stimuli, to rule out the use of alternate solution strategies. For primacy (see main text), AAAA and ABBA should be classified similarly to new $n = 2$ CFG and FSG patterns, respectively, presented during the same probe sessions (Methods). For recency (see main text), BBBB and BAAB should be classified similarly to new $n = 2$ CFG and FSG patterns, respectively. If starlings are listening for the presence of a B/A motif transition (see text), then the d' value comparing BAAB and ABBA to AAAA and BBBB should be similar to that for new $n = 2$ CFG and FSG patterns. d' for the new $n = 2$ CFG and FSG probe stimuli (dark blue) was significantly higher than that for all three control comparisons (light blue; asterisk indicates $P < 0.05$ for all cases, paired t -test).

practice, however, the stimulus sets used to test such claims must be finite. Thus, the theoretical possibility remains that a finite-state grammar, however heavily contrived, may account for the observed behaviour (see Supplementary Information). Of course, theoretical difficulties in proving the use of context-free rather than finite-state grammars extend to studies of grammatical competence in humans as well, and therefore call into question the falsifiability of claims regarding CFGs in humans compared to non-humans.

Although uniquely human syntactic processing capabilities, if any, may reflect more complex context-free grammars or higher levels in the Chomsky grammatical hierarchy, it may prove more useful to consider species differences as quantitative rather than qualitative distinctions in cognitive mechanisms. Such mechanisms (for example, memory capacity) need not map precisely onto strict formal grammars and automata theories. There might be no single property or processing capacity that marks the many ways in which the complexity and detail of human language differs from non-human communication systems¹³. Future studies can gauge the extent of the recursive syntactic abilities demonstrated here, by examining the processing of right-embedded structures more common in human languages (and more easily understood), and the interface between generalized syntactic and semantic knowledge.

METHODS

Baseline training. Motifs can be classified into four spectro-temporally distinct categories: whistles, warbles, rattles and high-frequency motifs⁶ (Supplementary Information). We used eight 'rattle' and eight 'warble' motifs from the repertoire of a single male starling (sets a_i and b_i , respectively, $i = 1-8$) as the vocabulary for two distinct grammars (Fig. 2a, b). One grammar defined sequences of the form A^2B^2 , and the other defined sequences of the form $(AB)^2$. For example, the explicit sequence of sound patterns rattle₁-rattle₁-warble₁-warble₁ is defined by A^2B^2 , but rattle₁-warble₁-rattle₁-warble₁, using the same four elements, is defined by $(AB)^2$. Because the song stimuli were created from a common vocabulary, only the patterning of motifs within each sequence varied between the classes defined by each grammar.

For each of the CFG (A^2B^2) and FSG ($(AB)^2$) grammars, we generated all possible motif sequences consisting of four elements. From each of these complete 'languages', we randomly selected three subsets of eight sequences such that within each subset, each motif appeared exactly once in each possible

position, and no motif appeared twice in the same sequence. We chose one subset from each of the two grammars (sixteen sequences, eight per grammar) as the stimuli for initial operant classification training (Fig. 2 and Supplementary Information).

We trained starlings to classify sequences defined by the FSG and CFG using well-established operant procedures⁹ (Supplementary Information). Starlings learned to respond regularly to the sequences defined by one grammar (S^+ stimuli) and to withhold responses to sequences defined by the other grammar (S^- stimuli). For half of the birds the CFG sequences served as S^+ stimuli and the FSG sequences as S^- stimuli, and for the other half these associations were reversed. Both baseline training groups (CFG sequences as S^+ , $n = 5$; or CFG sequences as S^- , $n = 4$) learned at roughly equivalent rates. (There was no significant difference between groups in the mean number of trials required to reach criterion, $P = 0.46$, Mann–Whitney U -test.)

d' measure of classification. We used a learning criterion of $d' > 1.0$ and a lower bound of 95% confidence interval around d' above zero for five consecutive blocks (100 trials per block). The measure of d' indexes the subject's sensitivity to differentiating between the two classes of patterns presented as stimuli, uncontaminated by response bias (Supplementary Information).

Probe-session reinforcement contingencies. During probe sessions we presented familiar CFG and FSG sequences on 80% of all trials, and various new test ('probe') stimuli on the remaining 20% of trials. We reinforced responses to the familiar stimuli in the normal manner (but at a reduced rate), and randomly reinforced all responses to the probe stimuli, regardless of accuracy, with equal rates of food and a short time-out (see Supplementary Information). As probe stimuli, we presented new $A^n B^n$ and $(AB)^n$ patterns ($n = 2, 3$ or 4), agrammatical patterns of the form $AAAA$, $BBBB$, $ABBA$, $BAAB$, and patterns defined by the FSG A^*B^* .

Higher-order stimuli. We used the CFG and FSG grammars to randomly generate 16 motif sequences with $n = 3$ (that is, A^3B^3 and $(AB)^3$, eight per grammar) and 16 sequences with $n = 4$ (eight per grammar). As for $n = 2$, for each set of eight exemplars from a given grammar at a given order, each motif appeared in every position once and only once, and the same motif never appeared more than once in the same exemplar sequence. We then presented the $n = 3$ and $n = 4$ sequences as probe stimuli (in separate sessions for each n) while birds continued classification of the $n = 2$ baseline training stimuli.

Received 21 October; accepted 27 February 2006.

1. Hockett, C. F. The origin of speech. *Sci. Am.* **203**, 89–96 (1960).
2. Chomsky, N. *Syntactic Structures* (Mouton, The Hague, 1957).
3. Hopcroft, J. & Ullman, J. *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, Reading, Massachusetts, 1979).
4. Hauser, M. D., Chomsky, N. & Fitch, W. T. The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579 (2002).
5. Fitch, W. T., Hauser, M. D. & Chomsky, N. The evolution of the language faculty: clarifications and implications. *Cognition* **97**, 179–210 (2005).
6. Eens, M. Understanding the complex song of the European starling: An integrated approach. *Adv. Study Behav.* **26**, 355–434 (1997).
7. Gentner, T. Q. & Hulse, S. H. Perceptual mechanisms for individual vocal recognition in European starlings, *Sturnus vulgaris*. *Anim. Behav.* **56**, 579–594 (1998).
8. Gentner, T. Q. & Hulse, S. H. Perceptual classification based on the component structure of song in European starlings. *J. Acoust. Soc. Am.* **107**, 3369–3381 (2000).
9. Gentner, T. Q. & Margoliash, D. Neuronal populations and single cells representing learned auditory objects. *Nature* **424**, 669–674 (2003).
10. Hauser, M. D., Newport, E. L. & Aslin, R. N. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition* **78**, B53–B64 (2001).
11. Hauser, M. D., Weiss, D. & Marcus, G. Rule learning by cotton-top tamarins. *Cognition* **86**, B15–B22 (2002).
12. Fitch, W. T. & Hauser, M. D. Computational constraints on syntactic processing in a nonhuman primate. *Science* **303**, 377–380 (2004).
13. Pinker, S. & Jackendoff, R. The faculty of language: what's special about it? *Cognition* **95**, 201–236 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Brawn for help in conducting some of these experiments, and A. Henly, P. Visser and L. Kay for comments on an earlier draft. This research was supported by an NIH grant to D.M.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.Q.G. (tgentner@ucsd.edu).

LETTERS

TMP21 is a presenilin complex component that modulates γ -secretase but not ϵ -secretase activity

Fusheng Chen^{1*}, Hiroshi Hasegawa^{1*†}, Gerold Schmitt-Ulms^{1*}, Toshitaka Kawai^{1†}, Christopher Bohm¹, Taiichi Katayama¹, Yongjun Gu¹, Nobuo Sanjo^{1†}, Michael Glista¹, Ekaterina Rogaeva¹, Yosuke Wakutani¹, Raphaëlle Pardossi-Piquard¹, Xueying Ruan¹, Anurag Tandon¹, Frédéric Checler², Philippe Marambaud³, Kirk Hansen⁴, David Westaway¹, Peter St George-Hyslop^{1,5} & Paul Fraser¹

The presenilin proteins (PS1 and PS2)^{1,2} and their interacting partners nicastrin³, aph-1 (refs 4, 5) and pen-2 (ref. 5) form a series of high-molecular-mass, membrane-bound protein complexes^{6–8} that are necessary for γ -secretase and ϵ -secretase cleavage of selected type 1 transmembrane proteins, including the amyloid precursor protein⁹, Notch¹⁰ and cadherins¹¹. Modest cleavage activity can be generated by reconstituting these four proteins in yeast and *Spodoptera frugiperda* (sf9) cells^{12–14}. However, a critical but unanswered question about the biology of the presenilin complexes is how their activity is modulated in terms of substrate specificity and/or relative activities at the γ and ϵ sites. A corollary to this question is whether additional proteins in the presenilin complexes might subsume these putative regulatory functions. The hypothesis that additional proteins might exist in the presenilin complexes is supported by the fact that enzymatically active complexes have a mass that is much greater than predicted for a 1:1:1:1 stoichiometric complex (at least 650 kDa observed, compared with about 220 kDa predicted)^{6–8}. To address these questions we undertook a search for presenilin-interacting proteins that differentially affected γ - and ϵ -site cleavage events. Here we report that TMP21, a member of the p24 cargo protein family, is a component of presenilin complexes and differentially regulates γ -secretase cleavage without affecting ϵ -secretase activity.

To isolate additional presenilin complex components we used an affinity-purified polyclonal antibody (Ab4) directed against the amino terminus of PS1 to immunoprecipitate PS1 complexes from CHAPSO-solubilized membranes of wild-type blastocyst-derived cells expressing PS1 and PS2, and from similar membranes of PS1^{-/-}/PS2^{-/-} double-knockout blastocyst-derived cells (which served as a negative control). The complexes immunopurified from wild-type cells had a mass of at least 650 kDa and possessed γ -secretase enzymatic activity⁶. The immunoprecipitates from both wild-type and PS1/PS2 double-knockout cells were subjected to in-solution tryptic digestion, and the protein components were then identified by mass spectrometry.

All four known constituents of the PS1 complex (nicastrin³, aph-1 (refs 4, 5), pen-2 (ref. 5) and PS1 (ref. 1)) were present only in the co-precipitates from wild-type cells. Among the few other proteins that were uniquely present only in the immunoprecipitates from wild-type presenilin-expressing cells but not in those from the PS1^{-/-}/PS2^{-/-} cells, the strongest identification was made

for TMP21 (accession number Q9D1D4) based on three unique tryptic peptides (LKPLEVELR, IPDQLVILDMK and RLEDLSE-SIVNDFAYMK) covering 17.5% of the full protein length. TMP21, a 219-amino-acid type 1 transmembrane protein, is a member of the p24 cargo-protein family¹⁵ that is involved in protein transport and quality control in the endoplasmic reticulum and Golgi¹⁶. TMP21 protein also resides at the plasma membrane¹⁵ (which is one of the principal subcellular locations for γ - and ϵ -cleavage of many substrates). The gene encoding TMP21 is located on chromosome 14 in a highly conserved cluster of genes that maps close to PS1 itself^{1,17}. As might be predicted, recent bioinformatic analyses reveal that TMP21, PS1 and amyloid precursor protein (APP) display dynamic patterns of co-transcription¹⁸.

The authenticity of the interaction between TMP21 and the presenilin complex was confirmed by showing that endogenous

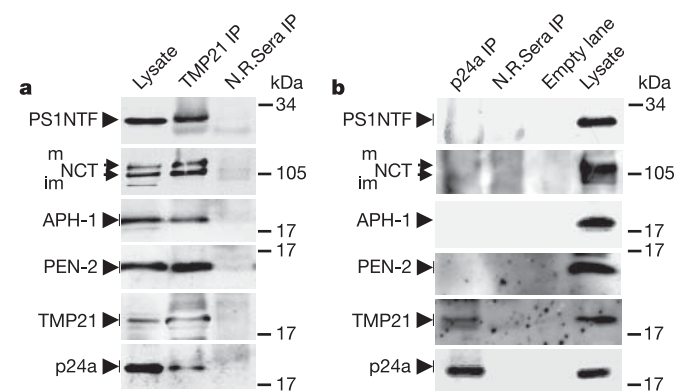


Figure 1 | TMP21 co-precipitates with other presenilin complex components. **a**, Both endogenous and exogenous Flag-tagged TMP21 reciprocally co-precipitate known members of the presenilin complex from HEK-293 cells, SHSY-5Y cells and mouse brain (Supplementary Fig. 1a). TMP21 also interacts with the p24 cargo protein p24a, but as a component of a separate complex. N.R.Sera IP, non-reactive serum immunoprecipitation control; mNCT, mature nicastrin; imNCT, immature nicastrin. Flag-TMP21, Flag-epitope-tagged TMP21, which was used in some experiments as indicated. **b**, p24a, another member of the p24 cargo protein family, does not interact with any members of the presenilin complex but does co-precipitate TMP21.

¹Departments of Medicine, Laboratory Medicine and Pathobiology, and Medical Biophysics, and Centre for Research in Neurodegenerative Diseases, University of Toronto, Toronto, Ontario, Canada M5S 3H2. ²Institut de Pharmacologie Moléculaire et Cellulaire, UMR6097CNRS/UNSA, Equipe labellisée FRM, 06560 Valbonne, France. ³Department of Pathology, Albert Einstein College of Medicine, Bronx, New York, New York 10461, USA, and The Litwin-Zucker Research Center, North Shore-LIJ Institute for Medical Research, Manhasset, New York 11030, USA. ⁴Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143, USA, and Proteomics Core, University Health Sciences Center, Aurora, Colorado 80045, USA. ⁵Department of Medicine (Division of Neurology), Toronto Western Hospital Research Institute, University Health Network, Toronto, Ontario, Canada M5G 1Z5. [†]Present addresses: Shiga University, Otsu, Japan (H.H.); Brain and Heart Centre, Hyogo, Japan (T.K.); Tokyo Medical and Dental University, Tokyo, Japan (N.S.).

*These authors contributed equally to this work.

TMP21 from mouse brain, neuron-like SHSY-5Y cells and human embryonic kidney (HEK-293) cells could be co-immunoprecipitated with endogenous nicastrin, aph-1, pen-2 and PS1 (Fig. 1a, and Supplementary Fig. 1a) and that it had an overlapping size distribution with presenilin complex components in high-molecular-mass (more than 650-kDa) fractions on glycerol velocity gradients and on two-dimensional Blue Native gel chromatography (Supplementary Fig. 1b, c). Furthermore, in the absence of PS1 and PS2, or in the absence of pen-2, TMP21 was destabilized from the complexes with a molecular mass of more than 650 kDa, co-localizing predominantly with an approximately 150-kDa nicastrin–aph-1 complex and with an approximately 440-kDa complex containing the remaining presenilin complex components (Supplementary Fig. 1c). TMP21 could be surface biotinylated, as can nicastrin (Supplementary Fig. 2), and co-localized with presenilin complex components in biochemical fractionation and immunofluorescence studies in the endoplasmic reticulum, Golgi and cell surface (not shown). In contrast, p24a, another member of the p24 cargo-protein family with 48% amino-acid sequence similarity to TMP21 ($E = 7 \times 10^{-9}$, where E is the expected number statistic for high scoring sequence alignment pairs) forms heteromeric complexes with TMP21 in the endoplasmic reticulum and Golgi^{15,16}, but does not co-precipitate with presenilin complex components (Fig. 1b).

To assess the functional consequences of the interaction between TMP21 and presenilin complexes, we next investigated the effects of modulating expression of TMP21. Transient overexpression of TMP21 had no discernible effect on the abundance of the presenilin complex components, on the abundance of p24 proteins such as p24a, on the subcellular distribution of these proteins (as assessed by

both biochemical fractionation and immunocytochemistry data (not shown)), or on A β production in either whole HEK-293 cells or in cell-free γ -secretase assays (Supplementary Fig. 3a).

In contrast, when TMP21 expression was suppressed by small interfering RNAs (siRNAs) there was an increase in the production of both A β 40 and A β 42. The increase in A β production was observed in whole HEK-293 cells overexpressing APPswedish (A β was $189 \pm 20.70\%$ (mean $\pm$ s.e.m.) of control, $n = 5$, $P < 0.005$; Fig. 2a, left panel, and Fig. 2b), in HEK-293 cells overexpressing wild-type APP ($216 \pm 20.07\%$ of control, $n = 5$, $P = 0.005$; not shown), in native HEK-293 cells expressing endogenous APP ($181 \pm 17.18\%$ of control, $n = 5$, $P = 0.005$; not shown) and in neuron-like SHSY-5Y cells ($219 \pm 10\%$, $n = 3$, $P < 0.01$; Supplementary Fig. 3b). Similar increases in A β production were also observed in cell-free γ -secretase assays of TMP21-deficient presenilin complexes with either endogenous 'pre-docked' C100-APP substrate (Fig. 2a, left panel, and Fig. 2b), or with an exogenous recombinant C100-APP substrate (Fig. 2a, left panel, and Fig. 2b). This increase in γ -secretase activity was not accompanied by either changes in the levels of endogenous PS1, nicastrin, aph-1, pen-2, N'-O'-glycosylated APP holoprotein or changes in α -secretase or β -secretase activity (the levels of secreted N-terminal APP ectodomain fragments in the conditioned media were unaltered; Fig. 2a, left panel). These results indicate that although TMP21 might be a component of presenilin complexes, it is not essential for the assembly of the presenilin complexes. This result also indicates that TMP21 modulates γ -secretase activity by a method other than by simply altering presenilin complex assembly and stability.

The increase in A β secretion after suppression of TMP21 was

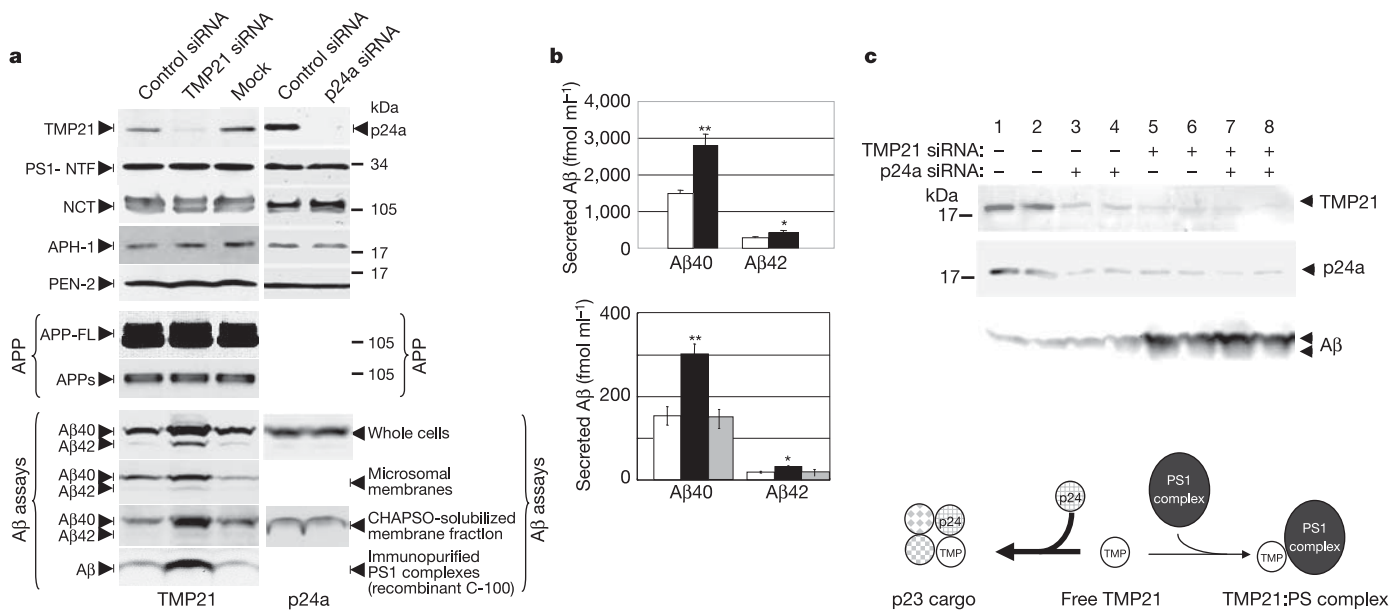


Figure 2 | Knockdown of TMP21 increases A β production. **a**, Suppression of TMP21 by siRNA (left panel, middle column) causes increased γ -secretase activity and increased production of A β from whole cells, microsomal membranes, CHAPSO-solubilized membranes or immunopurified PS1 complexes. In contrast, suppression of another p24 cargo protein, p24a, by siRNA has no effect on A β production (right panel, right column). Control siRNA, scrambled nonsense siRNA oligonucleotides; mock, no siRNA oligonucleotides; APPs, secreted N-terminal ectodomain fragments of APP. **b**, Quantification of A β secretion by whole-cell (upper panel) or by cell-free (lower panel) γ -secretase assays after suppression of TMP21 or p24a by siRNA. Black bars, TMP21 siRNA; white bars, control scrambled siRNA; grey bars, p24a siRNA. Asterisk, $P < 0.01$; two asterisks, $P < 0.001$. Error bars show s.e.m. **c**, For equivalent levels of p24a suppression (mediated by p24a siRNA only, by TMP21 siRNA or by both p24a siRNA and TMP21 siRNA) the suppression of p24a-associated TMP21 (by p24a siRNA only;

lanes 3 and 4) had no effect on A β , whereas the subsequent additional suppression of the remaining free TMP21 (by p24a siRNA and TMP21 siRNA; lanes 7 and 8) causes an increase in A β . This increase in A β production in the double (p24a and TMP21) siRNA-suppressed cells (lanes 7 and 8) was equivalent to the increase in A β production induced by siRNA suppression of TMP21 alone (lanes 5 and 6). The figure is not directly intended to dissect the relative compartment sizes of free TMP21 and p24a-bound TMP21. However, by comparing the decrease in TMP21 signal between the p24a siRNA (lanes 3 and 4, which suppresses mostly p24a-bound TMP21) and the TMP21 siRNA (lanes 5 and 6, which suppresses both free TMP21 and p24a-bound TMP21), one can infer that in the presence of equivalent levels of p24a, the TMP21 siRNA induced an approximately 10–20% decrease in TMP21 signal intensity. This decrease represents the decrease in free TMP21.

specific to TMP21. Neither transient overexpression (Supplementary Fig. 3a) nor siRNA-mediated suppression of p24a (Fig. 2a, right panel, and Fig. 2b) had any effect on either the abundance of the presenilin complex proteins or on A β production, regardless of whether whole cells or CHAPSO-solubilized microsomal membranes were used (Supplementary Fig. 4). However, as expected, p24a suppression did decrease TMP21 levels in whole cells (Fig. 2c). This apparent paradox, in which there is no change in A β production when decreases in TMP21 are caused by decreases in p24a but where there are profound changes in A β production when TMP21 itself is suppressed, can be explained by the existence of two pools of TMP21 (Fig. 2c, bottom panel). The major pool is stabilized by p24a but has no direct role in A β production. In contrast, the second and smaller pool of TMP21 interacts tightly with presenilin complexes and regulates A β production. The existence of these two pools of TMP21 is shown by the following experiments. The decrease in p24a-associated TMP21 (through p24a suppression) had no effect on A β levels. However, there was a significant increase in A β production when the residual 'non-p24a-associated' TMP21 in these p24a siRNA-treated cells was subsequently suppressed by TMP21 siRNA. This increase in A β production in the double (p24a and TMP21) siRNA-suppressed cells was equivalent to the increase in A β production induced by the siRNA suppression of TMP21 alone (Fig. 2c).

TMP21 could regulate the activity of the presenilin complex through at least three mechanisms. First, it is conceivable that TMP21 might be a competing substrate for presenilin-mediated endoproteolysis. This hypothesis can be excluded by the following facts: no N-terminal secreted products of TMP21 could be detected in conditioned medium (data not shown); no C-terminal TMP21 fragments corresponding to a putative γ -site cleavage product could be detected in cell lysates; and no carboxy-terminal precursors for γ -secretase cleavage of TMP21 equivalent to α - or β -secretase stubs were detectable even in cells treated with potent γ -secretase inhibitors (such as compound E) or in cells deficient in both PS1 and PS2 (Supplementary Fig. 5). Second, as a member of the p24 cargo protein family, it is conceivable that TMP21 might modulate the maturation and/or subcellular trafficking of the presenilin complex and its substrates. However, pulse-chase and surface biotinylation analyses, after either overexpression or underexpression of TMP21, revealed no detectable changes in the glycosylation, maturation, abundance or temporal patterns of trafficking of the APP substrate or of the presenilin complex components to the cell surface (Supplementary Fig. 6a, b). Thus, suppression of TMP21 had no effect on the abundance of APP and nicastrin at the cell surface as measured by surface biotinylation (Supplementary Fig. 6b), on the patterns of glycosylation of APP and nicastrin (Fig. 2a), or on the patterns of endoproteolysis of PS1 (Fig. 2a).

The third potential mechanism is that, in addition to its general role as a p24 cargo protein, TMP21 might also act as a direct

modulator of the presenilin complex itself (Fig. 2c, bottom panel). To resolve whether TMP21 had a direct effect on the function of presenilin complexes that was independent of any role that TMP21 might have on trafficking. We investigated the effects of adding exogenous Flag-tagged TMP21 to immunopurified, TMP21-deficient presenilin complexes. These cell-free complementation experiments, in which issues relating to trafficking can be excluded, revealed that A β production reverted to normal when exogenous TMP21 was added to the immunopurified TMP21-deficient PS1 complexes (Fig. 3). However, A β production remained elevated in the 'mock-complemented' TMP21-deficient PS1 complexes (Fig. 3).

Surprisingly, and in notable contrast to its effect on γ -secretase activity, the suppression of TMP21 by siRNA had no discernible effect on ϵ -secretase activity as measured by the production of the amyloid intracellular domain (AICD), the Notch intracellular domain (NICD) or the Cadherin intracellular domain (CICD) (Fig. 4). This result was robust regardless of whether ϵ -secretase activity was assayed in whole cells with endogenous substrate or in

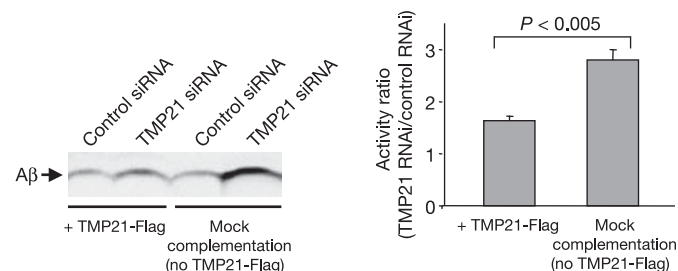


Figure 3 | Complementation of TMP21-deficient presenilin complexes (TMP21 siRNA) by exogenous immunopurified TMP21 (+TMP21-Flag) reverts γ -secretase activity towards levels observed in wild-type control complexes (Control siRNA). In contrast, A β production remains elevated in TMP21-deficient complexes treated with anti-Flag immunoprecipitates lacking TMP21 (+ control, no TMP21-Flag). Error bars show s.e.m.

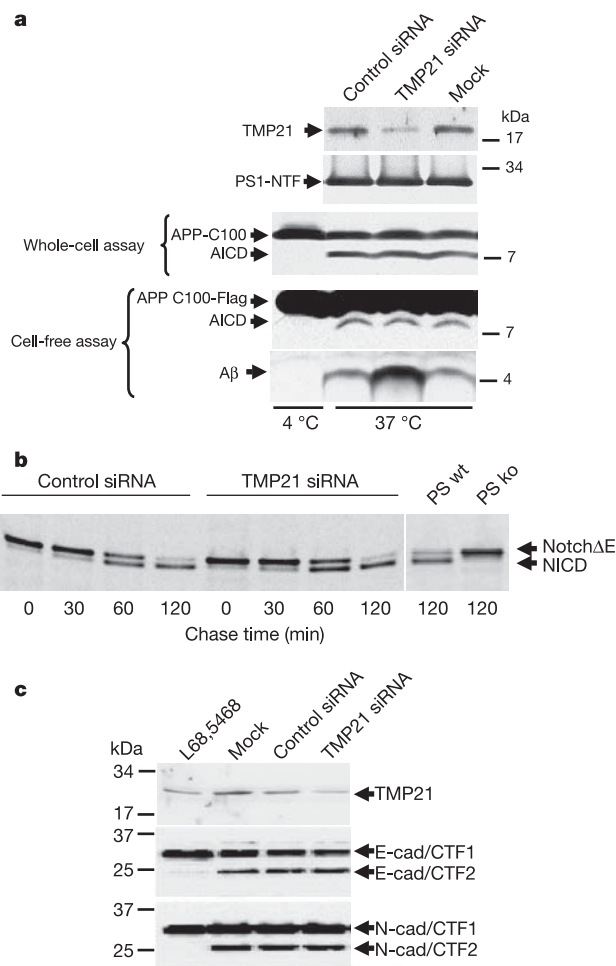


Figure 4 | The ϵ -secretase cleavage site is not affected by knockdown of TMP21. **a**, TMP21 siRNA suppression does not affect either the cleavage of endogenous APP by ϵ -secretase to generate AICD by endogenous PS1 complexes in whole-cell assays or the ϵ -secretase-mediated cleavage of recombinant C100-APP (equivalent to the natural β -secretase-generated substrate) by immunopurified PS1 complexes in cell-free assays. **b**, TMP21 siRNA has no effect on either the kinetics or the amount of cleavage of the Notch substrate (Notch Δ E) by ϵ -secretase to generate NICD in whole cells. **c**, TMP21 siRNA suppression has no effect on ϵ -cleavage of either E-cadherin (E-cad) or N-cadherin substrates (CAD/CTF1) to generate CAD/CTF2 C-terminal products. In contrast, this cleavage can be specifically inhibited by the γ / ϵ -secretase inhibitor L685,458.

purified complexes with a recombinant substrate (Fig. 4a). This result also strongly supports the conclusion that the effect of TMP21 suppression is not due to a simple defect in vesicular trafficking, but is due to a specific effect on one aspect of the function of PS1 complexes. Nevertheless, this result is surprising because previous studies have shown that suppression of the *Caenorhabditis elegans* homologue of p24a (*sel-9*) restores the signalling activity of mutant *lin-12* or *glp-1* proteins by permitting trafficking of the mutant proteins to the cell surface¹⁹. However, this latter effect probably reflects the separate general role of p24 cargo proteins in the quality control of mutant proteins in the secretory pathway (that is, the suppression of *sel-9* expression in cells expressing mutant *lin-12*/*glp-1* abrogates the normal quality-control function of *sel-9*, which would normally inhibit the transport of the mutant *lin-12* and *glp-1* to the cell surface)¹⁹.

Taken together, our data indicate that, in addition to its function in protein transport and quality control within the secretory pathway, TMP21 might also have a specific role as a modulator of presenilin-dependent γ -site cleavage. This newly discovered property is highly specific to the γ -site cleavage, and TMP21 does not modulate cleavage at the ϵ site. Moreover, this effect is restricted to TMP21 and is not a general property of other p24 cargo proteins.

The precise mechanism by which TMP21 acts as a specific negative modulator of γ -site cleavage is not immediately apparent. However, the concept of a multimeric protein complex that releases constitutive inhibitory subunits only in response to specific stimuli is analogous to the NF- κ B/I κ B complex²⁰. Our results also indicate three further hypotheses. First, there may be additional proteins that might regulate other aspects of presenilin complex function (for example, potential modulators of ϵ -site cleavage or of substrate specificity). This conclusion is indirectly supported by recent genetic studies in *Drosophila melanogaster* indicating that the protein crumbs may be a negative modulator of presenilin-dependent Notch signalling²¹. Second, the TMP21-binding and modulatory domain is likely to be on the surface of the presenilin complex because complementation was achieved without requiring disaggregation and reassembly of the complex. Finally, in addition to supporting ligand-dependent signalling through ϵ -site cleavage of type 1 membrane proteins, it has also been suggested that the presenilins might also provide a 'garburator' mechanism to remove residual membrane-embedded stubs of membrane proteins^{22,23}. TMP21 provides a method of both preventing runaway intramembrane proteolysis and of coupling such a putative activity to other quality-control mechanisms mediated by TMP21 and its p24 cargo protein partners.

Our observation that γ - and ϵ -site cleavages are independently regulated strongly indicates that γ - and ϵ -site secretase cleavages are distinct but related properties of the presenilin complexes and are not simply the reflection of a loosely specified cleavage site by a single enzymatic activity. This conclusion is supported by the previous observations that some presenilin mutations²⁴ and some γ -secretase inhibitors²⁵ also have differential effects on γ - and ϵ -site cleavages. Understanding these mechanisms will probably be of use in the design of γ -site-specific inhibitors for the treatment of Alzheimer's disease. For instance, it will be of interest to determine whether the binding of small molecules or peptides to the luminal N terminus or to the cytosolic C terminus of TMP21 might regulate the effect of TMP21 on γ -site cleavage. If this is true, these ligands, or molecular mimics of TMP21 itself, might provide a way to manipulate γ -site cleavage (and thus A β production in patients with Alzheimer's disease) therapeutically, without altering ϵ -site cleavage (which is necessary for many physiological signal transduction mechanisms, including Notch signalling).

METHODS

Immunoaffinity purification of presenilin 1 complex and trypsinization. Membrane proteins were purified in parallel from wild-type and PS1^{-/-}/PS2^{-/-} blastocyst-derived cells extracted with buffer A (25 mM Hepes pH 7.4,

150 mM NaCl, 2 mM EDTA, protease inhibitor cocktail (Sigma)) containing 1% CHAPS. After ultracentrifugation, solubilized membrane fractions were subjected to immunoprecipitation of PS1 complexes with the use of Protein A-Sepharose that had been saturated and chemically cross-linked with an affinity-purified polyclonal rabbit antibody directed against the N terminus of PS-1. Captured proteins were pH-drop eluted, denatured in 6 M urea, reduced/alkylated and subjected to in-solution trypsinization. Proteolytic fragments were analysed with two-dimensional liquid chromatography coupled to electrospray tandem mass spectrometry as described²⁶.

Immunoprecipitations, co-immunoprecipitations and immunoblotting. Immunoprecipitations were performed in buffer A containing 1% Nonidet P-40 (ref. 27). Co-immunoprecipitations were performed in buffer A containing 1% digitonin or CHAPS as the solubilizing detergent²⁷. After immunoblotting on nitrocellulose, protein bands were detected by enhanced chemiluminescence (ECL; Amersham Biosciences). In some co-immunoprecipitation experiments, anti-rabbit or anti-mouse IgG (Fc) secondary antibodies (Pierce) were used for the detection of target proteins.

Two-dimensional gel electrophoresis of presenilin 1 complexes. Two-dimensional gel electrophoresis was performed as described previously⁶.

TMP21 and p24a complementary DNAs and transfections. Untagged TMP21, TMP21 tagged at the C terminus with the Flag epitope, and untagged p24a were generated by cloning the respective cDNAs into pcDNA4 or pcDNA6 expression vectors (Invitrogen). These constructs were transiently or stably transfected (Lipofectamine 2000) into native HEK-293 cells, into a derivative HEK-293 cell line that stably expresses the Swedish APP mutant (APP^{Swedish}), or into blastocyst-derived mouse cells²⁸.

Notch and cadherin assays. Notch and cadherin cleavage assays were performed as described previously^{27,29}.

RNA interference. siRNA-based knockdowns of target proteins were performed as described, in HEK-293 or blastocyst-derived cells³⁰. The oligonucleotide sequences are available in Supplementary Information.

A β assays. A β 40 and A β 42 levels were measured by ELISA and immunoprecipitation-western blotting as described previously, with the use of 12–24-h conditioned medium collected from native HEK-293 cells or from HEK-293 cells stably overexpressing APP (APP^{Swedish} or APP^{wt})^{3,24}.

Cell-free γ -secretase assays with endogenous APP were performed on microsome membranes from HEK-293 cells stably overexpressing APP^{Swedish} with or without TMP21 RNAi treatments as described previously²⁴. ϵ -stubs were detected by western blotting with anti-APP-CT antibody (Sigma). Cell-free γ -secretase assay was performed as described^{7,13}, with exogenous APP-C100 as the substrate was performed with recombinant APP-C100 peptides generated from a prokaryotic expression vector encoding the C-terminal 100 amino acids (596–695) of human APP (695-residue isoform) followed by Flag and His₆ sequences (C100-Flag-His₆). Solubilized γ -secretase was prepared as described^{7,13}.

When γ -secretase assays were performed from immunoprecipitations, polyclonal anti-PS1-NTF antibody (Ab4) was added to 1% CHAPS-solubilized membranes in buffer A (25 mM HEPES pH 7.4, 150 mM NaCl, 2 mM EDTA, with protease inhibitor cocktail (Sigma)). After incubation overnight at 4 °C, beads were washed three times in buffer A containing 0.5% CHAPS and once in HEPES buffer containing 25 mM HEPES, pH 7.0, 150 mM NaCl, 5 mM MgCl₂, 5 mM CaCl₂, protease inhibitor cocktail (Pierce), 0.25% CHAPS. The samples were then resuspended in HEPES buffer containing 0.25% CHAPS and 0.7 μ M C100-Flag, and incubated for 6 h at 4 °C or 37 °C. For complementation experiments, eluates from the TMP21-Flag immunoprecipitation (with anti-Flag M2-Agarose from mouse (Sigma)) from HEK-293 cells with or without TMP21-flag overexpression were added to the reaction system and incubated at 4 °C for 1 h followed by the addition of C100-Flag substrate (final concentration 0.7 μ M) and incubation at 37 °C for 6 h. Western blotting served for both the detection of PS1 and the quantitative analysis of A β and ϵ -stubs after the elution of SDS sample buffer from the immunoprecipitation slurry.

Received 15 January; accepted 21 February 2006.

- Sherrington, R. *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754–760 (1995).
- Rogaev, E. I. *et al.* Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* **376**, 775–778 (1995).
- Yu, G. *et al.* Nicastrin modulates presenilin-mediated *notch*/*glp1* signal transduction and β APP processing. *Nature* **407**, 48–54 (2000).
- Goutte, C., Tsunozaki, M., Hale, V. A. & Priess, J. R. APH-1 is a multipass membrane protein essential for the Notch signaling pathway in *Caenorhabditis elegans* embryos. *Proc. Natl Acad. Sci. USA* **99**, 775–779 (2002).

5. Francis, R. *et al.* aph-1 and pen-2 are required for Notch pathway signaling, γ -secretase cleavage of β APP, and presenilin protein accumulation. *Dev. Cell* **3**, 85–97 (2002).
6. Gu, Y. *et al.* The presenilin proteins are components of multiple membrane-bound complexes which have different biological activities. *J. Biol. Chem.* **279**, 31329–31336 (2004).
7. Li, Y.-M. *et al.* Presenilin 1 is linked with gamma-secretase activity in the detergent solubilized state. *Proc. Natl Acad. Sci. USA* **97**, 6138–6143 (2000).
8. Evin, G. *et al.* Transition-state analogue γ -secretase inhibitors stabilize a 900 kDa presenilin/nicastrin complex. *Biochemistry* **44**, 4332–4341 (2005).
9. De Strooper, B. *et al.* Deficiency of presenilin 1 inhibits the normal cleavage of amyloid precursor protein. *Nature* **391**, 387–390 (1998).
10. De Strooper, B. *et al.* A presenilin-1-dependent γ -secretase-like protease mediates release of Notch intracellular domain. *Nature* **398**, 518–522 (1999).
11. Marambaud, P. *et al.* A presenilin-1/ γ -secretase cleavage releases the E-cadherin intracellular domain and regulates disassembly of adherens junctions. *EMBO J.* **21**, 1948–1956 (2002).
12. Edbauer, D. *et al.* Reconstitution of γ -secretase activity. *Nature Cell Biol.* **5**, 486–488 (2003).
13. Kimberly, W. T. *et al.* Gamma-secretase is a membrane protein complex comprised of presenilin, nicastrin, Aph-1, and Pen-2. *Proc. Natl Acad. Sci. USA* **100**, 6382–6387 (2003).
14. Takasugi, N. *et al.* The role of presenilin cofactors in the γ -secretase complex. *Nature* **422**, 438–441 (2003).
15. Blum, R. *et al.* Tmp21 and p24A, two type I proteins enriched in pancreatic microsomal membranes, are members of a protein family involved in vesicular trafficking. *J. Biol. Chem.* **271**, 17183–17189 (1996).
16. Jenne, N., Frey, K., Brugger, B. & Wieland, F. T. Oligomeric state and stoichiometry of p24 proteins in the early secretory pathway. *J. Biol. Chem.* **277**, 46504–46511 (2002).
17. Trower, M. K. *et al.* Conservation of synteny between the genome of the pufferfish (*Fugu rubripes*) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease (AD3 locus). *Proc. Natl Acad. Sci. USA* **93**, 1366–1369 (1996).
18. Li, K. C., Liu, C. T., Sun, W., Yuan, S. & Yu, T. A system for enhancing genome-wide coexpression dynamics study. *Proc. Natl Acad. Sci. USA* **101**, 15561–15566 (2004).
19. Wen, C. & Greenwald, I. p24 proteins and quality control of LIN-12 and GLP-1 trafficking in *Caenorhabditis elegans*. *J. Cell Biol.* **145**, 1165–1175 (1999).
20. Luo, J. L., Kamata, H. & Karin, M. IKK/NF- κ B signaling: balancing life and death—a new approach to cancer therapy. *J. Clin. Invest.* **115**, 2625–2632 (2005).
21. Herranz, H., Stamatakis, E., Feiguin, F. & Milan, M. Self-refinement of Notch activity through the transmembrane protein Crumbs: modulation of γ -secretase activity. *EMBO Rep.* (in the press).
22. Yu, C. *et al.* Characterization of a presenilin-mediated amyloid precursor protein carboxyl-terminal fragment γ . Evidence for distinct mechanisms involved in γ -secretase processing of the APP and Notch1 transmembrane domains. *J. Biol. Chem.* **276**, 43756–43760 (2001).
23. Kopan, R. & Ilagan, M. X. Gamma-secretase: proteasome of the membrane? *Nature Rev. Mol. Cell Biol.* **5**, 499–504 (2004).
24. Chen, F. *et al.* Presenilin 1 mutations activate γ 42-secretase but reciprocally inhibit ϵ -secretase cleavage of amyloid precursor protein (APP) and S3-cleavage of notch. *J. Biol. Chem.* **277**, 36521–36526 (2002).
25. Petit, A. *et al.* New protease inhibitors prevent γ -secretase-mediated production of A β 40/42 without affecting Notch cleavage. *Nature Cell Biol.* **3**, 507–511 (2001).
26. Schmitt-Ulms, G. *et al.* Time-controlled transcardiac perfusion cross-linking for the study of protein interactions in complex tissues. *Nature Biotechnol.* **22**, 724–731 (2004).
27. Chen, F. *et al.* Nicastrin binds to membrane-tethered Notch. *Nature Cell Biol.* **3**, 751–754 (2001).
28. Donoviel, D. *et al.* Mice lacking both presenilin genes exhibit early embryonic patterning defects. *Genes Dev.* **13**, 2801–2810 (1999).
29. Marambaud, P. *et al.* A CBP binding transcriptional repressor produced by the PS1/ ϵ -cleavage of N-cadherin is inhibited by PS1 FAD mutations. *Cell* **114**, 635–645 (2003).
30. Hasegawa, H. *et al.* Both the sequence and length of the C terminus of PEN-2 are critical for intermolecular interactions and function of presenilin complexes. *J. Biol. Chem.* **279**, 46455–46463 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank I. Schulz for the anti-TMP21 antibody, and F. T. Wieland for the anti-p24a antibody. This study was supported by grants from the Canadian Institutes of Health Research, the Howard Hughes Medical Institute, Canadian Institutes of Health Research–Japan Science and Technology Trust, the Alzheimer Society of Ontario, the Ontario Research and Development Challenge Fund and the Canada Foundation for Innovation.

Author Contributions All authors contributed fully to the design, analysis and interpretation of the experiments described in this report.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.StG.-H. (p.hyslop@utoronto.ca).

A voltage-gated proton-selective channel lacking the pore domain

I. Scott Ramsey¹, Magdalene M. Moran¹†, Jayhong A. Chong¹† & David E. Clapham¹

Voltage changes across the cell membrane control the gating of many cation-selective ion channels. Conserved from bacteria to humans¹, the voltage-gated-ligand superfamily of ion channels are encoded as polypeptide chains of six transmembrane-spanning segments (S1–S6). S1–S4 functions as a self-contained voltage-sensing domain (VSD), in essence a positively charged lever that moves in response to voltage changes. The VSD ‘ligand’ transmits force via a linker to the S5–S6 pore domain ‘receptor’², thereby opening or closing the channel. The ascidian VSD protein Ci-VSP gates a phosphatase activity rather than a channel pore, indicating that VSDs function independently of ion channels³. Here we describe a mammalian VSD protein (H_v1) that lacks a discernible pore domain but is sufficient for expression of a voltage-sensitive proton-selective ion channel activity. H_v1 currents are activated at depolarizing voltages, sensitive to the transmembrane pH gradient, H^+ -selective, and Zn^{2+} -sensitive. Mutagenesis of H_v1 identified three arginine residues in S4 that regulate channel gating and two histidine residues that are required for extracellular inhibition of H_v1 by Zn^{2+} . H_v1 is expressed in immune tissues and manifests the characteristic properties of native proton conductances (G_{vH^+}). In phagocytic leukocytes⁴, G_{vH^+} are required to support the oxidative burst that underlies microbial killing by the innate immune system^{4,5}. The data presented here identify H_v1 as a long-sought voltage-gated H^+ channel and establish H_v1 as the founding member of a family of mammalian VSD proteins.

H_v1 was identified as a novel conserved human gene using bioinformatic searches based on known cation channels. The H_v1 gene is located on chromosome 12 at q24.11 and is encoded by seven exons, the first of which is predicted to be non-coding. H_v1 messenger RNA is encoded by multiple independent human expressed sequence tags that contain in-frame stop codons preceding the initiator ATG and 3' poly-A⁺ tails. The predicted H_v1 protein is 273 amino acids in length (31.7 kDa, pI = 6.62) and forms four transmembrane segments (S1–S4), according to hydrophathy analyses (Supplementary Fig. S1a, b). The overall transmembrane (TM) structure and placement of charged residues in hydrophobic domains is conserved among the vertebrate H_v1 orthologues (Supplementary Fig. S1a) and similar to both Ci-VSP (ref. 3) and the VSD of $K_v1.2$ (ref. 2). The amino-terminal ~100 amino acids in H_v1 are homologous to protein and lipid phosphatases, but unlike the catalytically active Ci-VSP, core active site residues required for phosphatase activity⁶ are not conserved in H_v1 .

Both the distribution of expressed sequence tags (UniGene, see <http://www.ncbi.nlm.nih.gov/UniGene/ESTProfileViewer.cgi?uglist=Hs.334637>) and high-throughput expression studies (GNF SymAtlas, see <http://symatlas.gnf.org/SymAtlas/>; search on MGC15619 protein) indicate that H_v1 mRNA is enriched in immune tissues such as lymph node, B-lymphocytes, monocytes and spleen. We found

marked lymph node expression of H_v1 mRNA on a human RNA dot blot (Supplementary Fig. S1d, spot F7). H_v1 protein expression was detected with an affinity-purified rabbit polyclonal anti-peptide antibody that recognizes H_v1 , green fluorescent protein (GFP)- H_v1 and H_v1 -haemagglutinin (HA) at their appropriate molecular weights in western blots of transfected HEK cells, but not mock-transfected controls (Supplementary Fig. S2b). The antibody recognized H_v1 in human immune tissues including lymph node (Supplementary Fig. S2a), and the human leukocyte cell lines HL-60 and Jurkat (Supplementary Fig. S2b, c).

In preliminary experiments, we observed large (>1 nA at +100 mV) voltage-dependent, outwardly rectifying whole-cell currents in transfected HEK cells bathed in standard Ringer's, low- Cl^- (where all but 8 mM Cl^- was substituted with NO_3^- , SCN^- or Br^-), or NMDG⁺(N-methyl-D-glucamine)·Cl (Na^+ , K^+ , Ca^{2+} and Mg^{2+} -free) solutions, suggesting that H_v1 currents were primarily permeable to H^+ . To study H^+ currents under better pH control, we used solutions with high H^+ buffering capacity (100 mM buffer used near its pK_a: MES, pH 5.5; Bis-Tris, pH 6.5; HEPES, pH 7.5) similar to those used previously to study native G_{vH^+} (refs 7, 8). In response to depolarizing voltage steps, cells transfected with H_v1 generated time-, voltage-, and pH-dependent currents with relatively slow activation kinetics and a rapidly decaying inward tail current (Fig. 1a–c). The kinetics of H_v1 activation ($\tau_{act} = 715 \pm 124$ ms) and deactivation ($\tau_{deact} = 65.0 \pm 12.1$ ms) were estimated from single exponential fits to the currents at +80 mV or –80 mV, respectively. Both τ_{act} and τ_{deact} are strongly voltage-dependent (Fig. 1e) and τ_{act} was noticeably variable between experiments, perhaps owing to small fluctuations in room temperature and the highly temperature-sensitive gating of H_v1 (Supplementary Fig. S3c, d).

In the absence of a H^+ gradient (the ratio of internal to external H^+ concentration $\rho H_{i/o}^+ = 1.0$), the voltage threshold (V_{thr}) for activation of outward currents (I_{step}) was 11 mV (Fig. 1d). An outwardly directed H^+ gradient, equivalent to net intracellular acidification ($\rho H_{i/o}^+ = 10$), caused V_{thr} to shift 40 mV negative (Fig. 1d). The reversal potential (E_{rev}) of open H_v1 channels determined from I_{tail} (Fig. 1e) was ~0 mV under symmetrical recording conditions ($\rho H_{i/o}^+ = 1$) and shifted >50 mV negative or positive in inward ($\rho H_{i/o}^+ = 0.1$) or outward ($\rho H_{i/o}^+ = 10$) H^+ gradients, respectively (Fig. 1f). A plot of E_{rev} versus $[H^+]$ was well fitted to a line with slope = 54.7 mV/log $[H^+]$ (data not shown), indicating that the H_v1 current was highly selective for H^+ . A plot of V_{thr} versus the calculated Nernst potential for H^+ was well fitted by the line $V_{thr} = 0.82E_{rev} + 13.8$ mV (Fig. 1g).

Although the outwardly rectifying I_{step} – V relation was linear at voltages ~10–70 mV positive to E_{rev} (Fig. 1d, Fig. 2a, c), we were surprised to find that the H_v1 I_{step} – V deviated from linearity and appeared to saturate at voltages ~100 mV positive to E_{rev} (Fig. 2a, c).

¹Howard Hughes Medical Institute, Department of Cardiology, Children's Hospital, Department of Neurobiology, Harvard Medical School, Enders 1309, 320 Longwood Avenue, Boston, Massachusetts 02115, USA. †Present address: Hydra Biosciences, 790 Memorial Drive, Cambridge, Massachusetts 02139, USA.

We infer that the apparent nonlinearity in I_{step} results from depletion of H_i^+ caused by rapid channel-mediated H^+ efflux that is not adequately compensated by delivery of H_i^+ from the buffer during the voltage step. Slow H_v1 kinetics near V_{thr} further hinder our ability to measure accurately the channels' voltage dependence

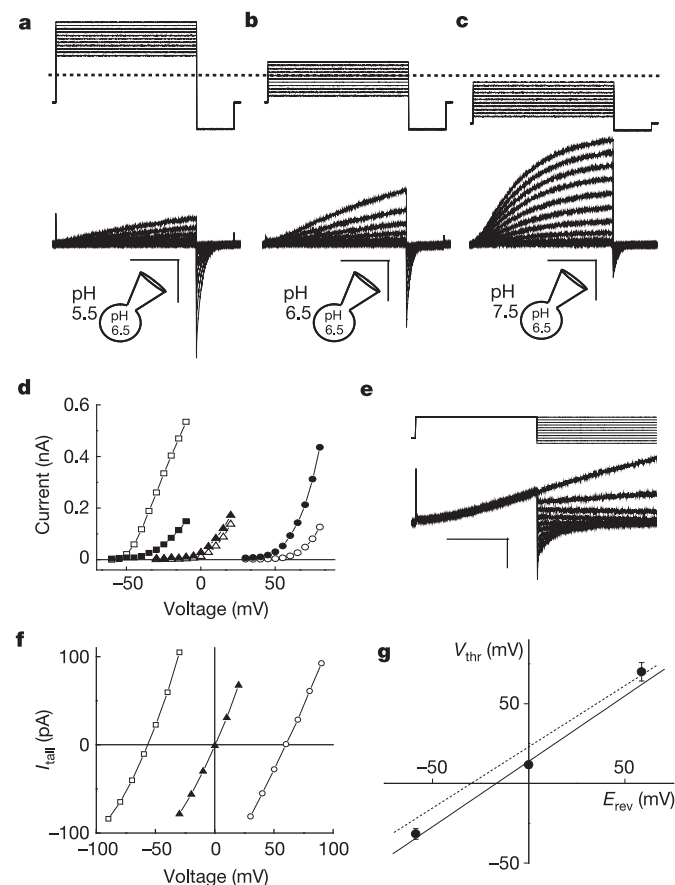


Figure 1 | Biophysical properties of expressed H_v1 currents. Depolarizing voltage steps (5 mV increments) were applied to an H_v1 -transfected HM1 cell (a–c) to elicit outward H^+ currents and deactivating inward tail currents (–80 mV). TMA5.5_o, TMA6.5_o or TMA7.5_o (TMA6.5_i for all) were used to impose pH gradients indicated in the diagrams. a, $\rho\text{H}_{i/o}^+ = 0.1$, $V_h = -40$ mV ($V_{\text{step}} = +30$ mV to $+80$ mV, scale bar 0.2 nA, 1 s); b, $\rho\text{H}_{i/o}^+ = 1$, $V_h = -40$ mV ($V_{\text{step}} = -30$ mV to $+20$ mV), scale bar 0.1 nA, 1 s; c, $\rho\text{H}_{i/o}^+ = 10$, $V_h = -70$ mV ($V_{\text{step}} = -60$ mV to -10 mV), scale bar 0.2 nA, 1 s. d, H_v1 currents at the end of the depolarizing step (I_{step} , open symbols) and the absolute value of I_{tail} (–80 mV, filled symbols) are plotted as a function of the step voltage. Data shown are from the cell shown in a–c. Circles, $\rho\text{H}_{i/o}^+ = 0.1$ ($V_{\text{thr}} = 70 \pm 5.8$ mV, $n = 3$); triangles, $\rho\text{H}_{i/o}^+ = 1.0$ ($V_{\text{thr}} = 11.7 \pm 2.1$ mV, $n = 6$); squares, $\rho\text{H}_{i/o}^+ = 10$ ($V_{\text{thr}} = -31.7 \pm 3.3$ mV, $n = 3$). e, Representative currents for E_{rev} measurement ($V_h = -40$ mV, $V_{\text{step}} = +40$ mV, $V_{\text{tail}} = -100$ mV to $+40$ mV). Small I_{step} (<200 pA) was chosen to minimize H_i^+ depletion (note constant outward current level). f, Monoexponential fits of I_{tail} were extrapolated to $t = 0$ and E_{rev} was estimated from the zero-current intercept of linear fits to the data. Open circles, TMA6.5_i, TMA5.5_o ($\rho\text{H}_{i/o}^+ = 0.1$); filled triangles, TMA6.5_i, TMA6.5_o ($\rho\text{H}_{i/o}^+ = 1$); open squares, TMA6.5_i, TMA7.5_o ($\rho\text{H}_{i/o}^+ = 10$). Average E_{rev} values: $E_{\text{rev}} = 53.3 \pm 1.4$ mV, $\rho\text{H}_{i/o}^+ = 0.1$, $n = 10$; $E_{\text{rev}} = 0.9 \pm 0.7$ mV, $\rho\text{H}_{i/o}^+ = 1.0$, $n = 14$; $E_{\text{rev}} = -56.5 \pm 2.6$ mV, $\rho\text{H}_{i/o}^+ = 10$, $n = 5$. Data represent mean $\pm$ s.e.m. from n experiments. g, V_{thr} is plotted against the Nernst potential for H^+ at 24 °C. The data are fitted to $V_{\text{thr}} = 0.82E_{\text{rev}} + 13.8$ mV (solid line). Data represent mean $\pm$ s.e.m. from $n = 3$ –7 experiments. The V_{thr} versus E_{rev} relationship for native G_{vH^+} ($V_{\text{thr}} = 0.79E_{\text{rev}} + 23$ mV⁴, dotted line) is shown for comparison.

because extraordinarily long voltage pulses are required to reach true steady state. Erosion of the H^+ gradient during I_{step} and underestimation of I_{step} are both expected to flatten the $I_{\text{tail}}-V$ relation. We were nonetheless able to fit H_v1 $I_{\text{tail}}-V$ relations ($\rho\text{H}_{i/o}^+ = 10$) to a Boltzmann function in order to approximate the voltage-dependence of the H_v1 open channel probability (P_o , Fig. 2c). In an outward H^+ gradient, H_v1 channels appear to be half open at ~ 60 mV and to move the equivalent of about one electronic charge across the membrane during channel opening ($z\delta = 0.9$). The midpoint of the $I_{\text{step}}-V$ relation shifted >40 mV when the H^+ gradient was altered, consistent with the observed shift in V_{thr} (Fig. 1).

In an initial attempt to determine whether voltage-dependent gating in H_v1 involved basic residues in S4, as it does in other voltage-gated channels, we individually replaced each of three arginine residues with alanine. Representative traces from cells expressing H_v1 or the point mutant R205A are shown in Fig. 2a, b. Relative to H_v1 , two arginine mutants (R205A, R208A) exhibited dramatically faster activation and deactivation kinetics, whereas R211A exhibited only faster deactivation kinetics (Supplementary Fig. S3e–g). R205A exhibited a positive shift in the half-maximal voltage for channel activation ($V_{0.5}$) and a statistically significant decrease in the slope factor ($z\delta$) derived from Boltzmann fits to $I_{\text{tail}}-V$ (Fig. 2c).

Like G_{vH^+} , H_v1 current was rapidly and reversibly blocked by bath superfusion of ZnCl_2 in the absence of EGTA (Fig. 3a, e). We therefore mutated putatively extracellular residues that might coordinate metal ions to inhibit H_v1 currents. Zn^{2+} potentially inhibited

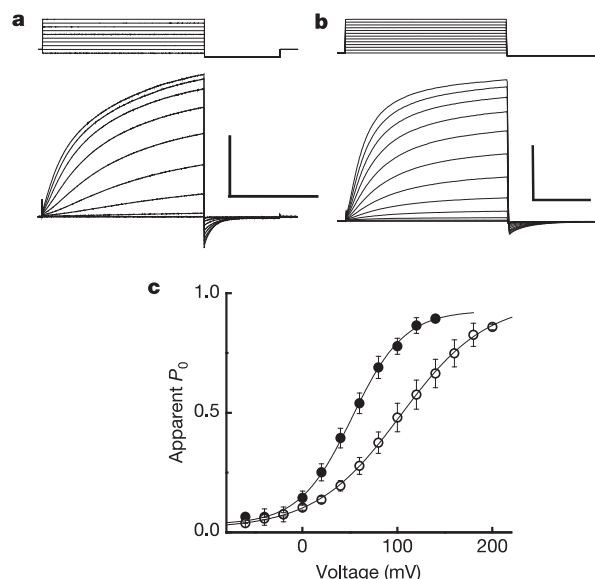


Figure 2 | H_v1 voltage-dependent gating. a, H_v1 currents (–60 mV to $+120$ mV, $V_h = -40$ mV, $\rho\text{H}_{i/o}^+ = 10$, Na6.5_i, Na7.5_o) in a representative cell. Scale bar 2 nA, 400 ms. Under symmetrical conditions (TMA6.5_i, $\rho\text{H}_{i/o}^+ = 1$), $\tau_{\text{act}} = 715 \pm 124$ ms ($+80$ mV), $n = 9$ and $\tau_{\text{deact}} = 65.0 \pm 12.1$ ms (–80 mV), $n = 7$. b, R205A (–60 mV to $+180$ mV, $V_h = -60$ mV, $\rho\text{H}_{i/o}^+ = 10$, Na6.5_i, Na7.5_o). Scale bar 2 nA, 10 ms. Note the 40-fold difference in timescale compared to a, c. The $I_{\text{tail}}-V$ relation (–80 mV, $\rho\text{H}_{i/o}^+ = 10$, Na6.5_i, Na7.5_o) was normalized to the maximum current obtained from a Boltzmann fit to the data for each cell expressing H_v1 (filled circles) or R205A (open circles) to estimate $P_o - V$. Data were fitted to a Boltzmann function (solid lines) and normalized to the extrapolated maximum current. Points represent mean $\pm$ s.e.m. of normalized data. A comparison of curve fits from individual experiments (H_v1 , $V_{0.5} = 58.0 \pm 5.6$ mV, $z\delta = 0.90 \pm 0.04$, $n = 6$; R205A, $V_{0.5} = 99.5 \pm 18.3$ mV, $z\delta = 0.57 \pm 0.03$, $n = 3$; $*P = 0.03$ by Student's non-paired t -test) indicates that significantly less effective charge is moved in R205A than in H_v1 during channel gating. Data represent mean $\pm$ s.e.m. from n experiments.

I_{Hv1} (half-maximal inhibitory concentration, $IC_{50} \approx 2 \mu M$) and slowed τ_{act} (Fig. 3a–c, data not shown). Apparent Zn^{2+} potency decreased 10- or 30-fold when we mutated H193 or H140 to alanine, respectively. Mutation of both residues (H140A/H193A) practically abolished Zn^{2+} sensitivity (Fig. 3d). All three histidine mutants exhibited the expected H^+ -selectivity and $\rho H_{i/o}$ -dependent shift in V_{thr} that are characteristic of H_v1 (data not shown).

The surprising finding of this study is that expression of a voltage-sensor domain, without its normally associated pore domain, is sufficient to reconstitute voltage-dependent proton channel activity. G_{vH+} have been recorded in a wide variety of cells from molluscs⁹ to mammals (reviewed in ref. 4). In mammalian phagocytes, G_{vH+} provides necessary charge compensation for the gp91^{phox} electron current that underlies Zn^{2+} -sensitive superoxide anion production^{4,5,10,11}. Loss of functional gp91^{phox} in humans causes X-linked chronic granulomatous disease, which is characterized by decreases in microbial killing by phagocytes^{12,13}.

The properties of expressed H_v1 are consistent with those of G_{vH+} (refs 4, 14, 15). H_v1 is H^+ -selective, exhibits a -40 mV shift in V_{thr} when pH_i is decreased one unit, is blocked by micromolar Zn^{2+} , and is highly temperature sensitive. H_v1 kinetics fall within the range

reported for G_{vH+} in a variety of mammalian cells⁴. Kinetic heterogeneity among native G_{vH+} could reflect the existence of unidentified H_v1 homologues, temperature differences during recording, post-translational modification of H_v1 , or second-messenger modulation of H^+ channels. The small endogenous G_{vH+} reported in HEK¹⁶ and CHO¹⁷ (but not COS-7¹⁸) cells appeared neither to facilitate nor interfere with expression of H_v1 current in this study. Slow activation kinetics and large H^+ currents in H_v1 -transfected cells will tend to exacerbate difficulties in measuring steady-state channel behaviour and in maintaining the imposed H^+ gradient during strongly depolarizing voltage pulses, and could explain why the voltage dependence of expressed H_v1 channels appears shallow relative to native currents⁴.

An accepted model for voltage-gated channels is one in which positively charged residues in S4 function as the primary voltage sensors^{19,20}. Charge-neutralizing mutagenesis of S4 arginine residues (R205A, R208A, R211A) demonstrates that these basic residues are intimately involved in H_v1 gating and may directly sense transmembrane voltage. Voltage-dependent gating in H_v1 is distinguished from other cation channels by its sensitivity to a H^+ gradient, indicating that further studies are required to clarify whether the mechanism of voltage sensing in H_v1 is similar to that in other voltage-gated channels. Starace and Bezanilla²¹ replaced an S4 arginine in *Shaker* with histidine (R362H) to create a H^+ -permeable channel at negative membrane potentials. In H_v1 , neutralizing S4 arginines or histidines in the extracellular linkers did not prevent the H^+ conductance, indicating that H^+ permeation does not require proton acceptors at these positions. Mutagenesis of putative H^+ acceptors may be of limited use in defining the conduction pathway given that H^+ conductance could be indirectly attenuated owing to allosteric effects. The mechanism of proton permeation may involve a voltage-dependent movement of S4 to align residues that create a 'proton wire', as in gramicidin²², or exposure of an H^+ acceptor that bridges two water-filled crevices²¹. The latter model unites both pore and wire features in that the proton acceptor can also be thought of as the binding site for H^+ within a pore that confers H^+ selectivity.

Polyvalent cations, particularly Zn^{2+} , are the only known blockers of native G_{vH+} (refs 4, 15). Zn^{2+} block of native proton current was predicted to involve two extracellularly accessible amino acids; apparent Zn^{2+} potency is modulated by both $[H^+]_o$ and membrane voltage¹⁵. Strikingly, we identified two histidine residues that appear to function coordinately to confer Zn^{2+} inhibition to H_v1 . In addition, Zn^{2+} slows τ_{act} , mirroring its effect on G_{vH+} (ref. 15) and further supporting the conclusion that H_v1 activity underlies native G_{vH+} .

Expression of H_v1 alone is sufficient to reconstitute most properties of G_{vH+} , making it unlikely that G_{vH+} is mediated by gp91^{phox} (refs 23, 24). Future studies are required to ascertain whether H_v1 physically associates with or is regulated by components of the NADPH oxidase complex, protein kinase C, or arachidonic acid^{25–27}, whether H_v1 channel activity is sufficient to control O_2^- secretion and reactive oxygen species production^{28,29}, and what factors account for differences between 'phagocyte' and 'epithelial' type G_{vH+} (ref. 4). Finally, future structural and biophysical studies should help to elucidate the proton permeation pathway within this voltage sensor domain protein.

Note added in proof: Okamura and colleagues also report that the mouse *Hv1* gene encodes a proton channel, which they call VSOP (voltage-sensor only protein)³⁰.

METHODS

Hv1 complementary DNA was amplified by polymerase chain reaction (PCR) from an expressed sequence tag clone (IMAGE 6424182) and subcloned into pQBI25-fC3 (QBiogene) to create *GFP-Hv1*, pcDNA3.1(–) for expression of non-tagged *Hv1*, or pBSIISK – (Stratagene) for *in vitro* mRNA transcription. Site-directed mutagenesis was used to create point mutations in *GFP-Hv1*: H140A, H193A, R205A, R208A, R211A, and the H140A/H193A double mutant;

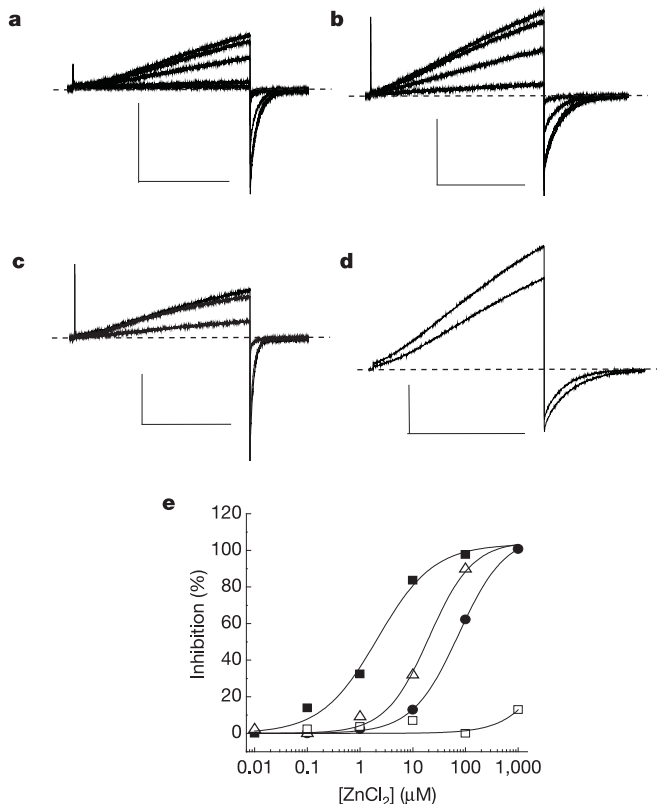


Figure 3 | Mutations in H_v1 reveal residues required for Zn^{2+} inhibition. Inhibition of I_{step} (+25 mV to +50 mV) in cells superfused with EGTA-free TMA6.5 or Na6.5 ($\rho H_{i/o} = 1$) was typically faster than the sampling interval (10 s) and washout was always complete. Average $IC_{50} = 2.2 \pm 0.6 \mu M$, $n_H = 1.0 \pm 0.2$, $n = 4$; $\rho H_{i/o} = 1$, +40 mV, TMA6.5. Data represent mean $\pm$ s.e.m. from n experiments. The applied voltages and $[Zn^{2+}]$ for the records shown (Na6.5, $\rho H_{i/o} = 1$) were: **a**, H_v1 , $V_{step} = +40/V_{tail} = -80$ mV, $[Zn^{2+}] = 0, 0.1, 1, 10, 100 \mu M$; **b**, H140A, +40/–60 mV, $[Zn^{2+}] = 0, 1, 10, 100 \mu M$; **c**, H193A, +30/–80 mV, $[Zn^{2+}] = 0, 10, 100 \mu M$; **d**, H140A/H193A, +50/–20 mV, $[Zn^{2+}] = 0, 1$ mM. Scale bars: 100 pA, 1 s. **e**, Representative Zn^{2+} concentration-response curves for H_v1 (filled squares, $IC_{50} = 1.9 \mu M$, $n_H = 0.9$), H193A (open triangles, $IC_{50} = 17.9 \mu M$, $n_H = 1.2$), H140A (filled circles, $IC_{50} = 74.3 \mu M$, $n_H = 1.0$), and H140A/H193A (open squares, maximum inhibition = $11.1 \pm 3.4\%$, $n = 3$).

H_v1 -HA was used for biochemical studies. A ~300 nucleotide antisense RNA probe labelled with α - 35 P-UTP was synthesized by *in vitro* transcription (Ambion) and hybridized (68 °C) to a human multiple tissue mRNA panel (MTE3, Clontech). Incorporated radioactivity was detected by a phosphorimager (Molecular Dynamics).

Currents reported here were recorded 18–36 h after transfection of $GFP-H_v1$ cDNA in the HEK-293 cell clone HM1. Recording solutions were similar to those described previously¹⁸ and contained 100 mM pH buffer (MES, Bis-Tris or HEPES) near its pK_a (5.5, 6.5 or 7.5, respectively) in tetramethylammonium methanesulphonate or NaCl adjusted to ~300 mOsm. H^+ gradients ($\rho_{H^+} = [H^+]_{intracellular}/[H^+]_{extracellular}$) were imposed by gravity-fed bath superfusion of differentially buffered solutions. Currents were recorded with an Axopatch 200A (Axon Instruments; 1 kHz filter) and digitized at 2 kHz (10 kHz for Arg mutants) using Clampex9 (Axon Instruments). Data were analysed using Clampfit9 (Axon Instruments) and Origin 6 (Microcal). Recordings were performed at 22–24 °C unless otherwise stated. Data shown and stated in text and legends represent mean $\pm$ s.e.m. for the indicated number of experiments.

Affinity-purified 4234 polyclonal rabbit antibodies (Chemicon) were raised against the keyhole limpet haemocyanin (KLH)-conjugated H_v1 peptide CSE-KEQEIERLNKL. Transfected cells were lysed in 1% Triton X-100 and crude membranes were subjected to SDS-polyacrylamide gel electrophoresis (PAGE) (NuPage, Invitrogen). Following transfer to nitrocellulose, blots were probed with the indicated concentration of 4234 followed by an anti-rabbit HRP-conjugated secondary antibody (Zymed). Immunoreactivity was detected by enhanced chemiluminescence (Pierce). A human immune tissue western blot (1525, ProSci) was probed with 4234 as described.

Received 5 February; accepted 8 March 2006.

Published online 22 March 2006.

- Yu, F. H. & Catterall, W. A. The VGL-channome: a protein superfamily specialized for electrical signaling and ionic homeostasis. *Sci. STKE* **2004**, re15 (2004).
- Long, S. B., Campbell, E. B. & Mackinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K^+ channel. *Science* **309**, 897–903 (2005).
- Murata, Y., Iwasaki, H., Sasaki, M., Inaba, K. & Okamura, Y. Phosphoinositide phosphatase activity coupled to an intrinsic voltage sensor. *Nature* **435**, 1239–1243 (2005).
- DeCoursey, T. E. Voltage-gated proton channels and other proton transfer pathways. *Physiol. Rev.* **83**, 475–579 (2003).
- DeCoursey, T. E., Morgan, D. & Cherny, V. V. The voltage dependence of NADPH oxidase reveals why phagocytes need proton channels. *Nature* **422**, 531–534 (2003).
- Goldberg, J. *et al.* Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature* **376**, 745–753 (1995).
- Schilling, T., Gratopp, A., DeCoursey, T. E. & Eder, C. Voltage-activated proton currents in human lymphocytes. *J. Physiol. (Lond.)* **545**, 93–105 (2002).
- DeCoursey, T. E., Cherny, V. V., DeCoursey, A. G., Xu, W. & Thomas, L. L. Interactions between NADPH oxidase-related proton and electron currents in human eosinophils. *J. Physiol. (Lond.)* **535**, 767–781 (2001).
- Thomas, R. C. & Meech, R. W. Hydrogen ion currents and intracellular pH in depolarized voltage-clamped snail neurones. *Nature* **299**, 826–828 (1982).
- Henderson, L. M., Chappell, J. B. & Jones, O. T. Superoxide generation by the electrogenic NADPH oxidase of human neutrophils is limited by the movement of a compensating charge. *Biochem. J.* **255**, 285–290 (1988).
- Rada, B. K., Geiszt, M., Kaldi, K., Timar, C. & Ligeti, E. Dual role of phagocytic NADPH oxidase in bacterial killing. *Blood* **104**, 2947–2953 (2004).
- Babior, B. M. NADPH oxidase. *Curr. Opin. Immunol.* **16**, 42–47 (2004).
- Smith, R. M. & Curnutte, J. T. Molecular basis of chronic granulomatous disease. *Blood* **77**, 673–686 (1991).
- DeCoursey, T. E. & Cherny, V. V. Temperature dependence of voltage-gated H^+ currents in human neutrophils, rat alveolar epithelial cells, and mammalian phagocytes. *J. Gen. Physiol.* **112**, 503–522 (1998).
- Cherny, V. V. & DeCoursey, T. E. pH-dependent inhibition of voltage-gated H^+ currents in rat alveolar epithelial cells by Zn^{2+} and other divalent cations. *J. Gen. Physiol.* **114**, 819–838 (1999).
- Eder, C. & DeCoursey, T. E. Voltage-gated proton channels in microglia. *Prog. Neurobiol.* **64**, 277–305 (2001).
- Cherny, V. V., Henderson, L. M. & DeCoursey, T. E. Proton and chloride currents in Chinese hamster ovary cells. *Membr. Cell Biol.* **11**, 337–347 (1997).
- Morgan, D., Cherny, V. V., Price, M. O., Dinanuer, M. C. & DeCoursey, T. E. Absence of proton channels in COS-7 cells expressing functional NADPH oxidase components. *J. Gen. Physiol.* **119**, 571–580 (2002).
- Bezanilla, F. The voltage sensor in voltage-dependent ion channels. *Physiol. Rev.* **80**, 555–592 (2000).
- Tombola, F., Pathak, M. M. & Isacoff, E. Y. How far will you go to sense voltage? *Neuron* **48**, 719–725 (2005).
- Starace, D. M. & Bezanilla, F. A proton pore in a potassium channel voltage sensor reveals a focused electric field. *Nature* **427**, 548–553 (2004).
- Pomes, R. & Roux, B. Molecular mechanism of H^+ conduction in the single-file water chain of the gramicidin channel. *Biophys. J.* **82**, 2304–2316 (2002).
- DeCoursey, T. E., Morgan, D. & Cherny, V. V. The gp91phox component of NADPH oxidase is not a voltage-gated proton channel. *J. Gen. Physiol.* **120**, 773–779 (2002).
- Henderson, L. M. & Meech, R. W. Evidence that the product of the human X-linked CGD gene, gp91-phox, is a voltage-gated H^+ pathway. *J. Gen. Physiol.* **114**, 771–786 (1999).
- Bokoch, G. M. Regulation of innate immunity by Rho GTPases. *Trends Cell Biol.* **15**, 163–171 (2005).
- DeCoursey, T. E., Cherny, V. V., Morgan, D., Katz, B. Z. & Dinanuer, M. C. The gp91phox component of NADPH oxidase is not the voltage-gated proton channel in phagocytes, but it helps. *J. Biol. Chem.* **276**, 36063–36066 (2001).
- Price, M. O. *et al.* Creation of a genetic system for analysis of the phagocyte respiratory burst: high-level reconstitution of the NADPH oxidase in a nonhematopoietic system. *Blood* **99**, 2653–2661 (2002).
- Bokoch, G. M. & Knaus, U. G. NADPH oxidases: not just for leukocytes anymore! *Trends Biochem. Sci.* **28**, 502–508 (2003).
- Qu, A. Y., Nanda, A., Curnutte, J. T. & Grinstein, S. Development of a H^+ -selective conductance during granulocytic differentiation of HL-60 cells. *Am. J. Physiol.* **266**, C1263–C1270 (1994).
- Sasaki, M., Takagi, M. & Okamura, Y. A novel protein with a voltage sensor domain is a voltage-gated proton channel. *Science* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. DeCoursey, C. Miller, R. MacKinnon and P. Bezanilla for comments on the manuscript, and K.-H. Lee for technical assistance. This work was supported by the Sandler Program for Asthma Research and the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.E.C. (dclapham@enders.tch.harvard.edu).

Targeting C-reactive protein for the treatment of cardiovascular disease

Mark B. Pepys¹, Gideon M. Hirschfield^{1†}, Glenys A. Tennent¹, J. Ruth Gallimore¹, Melvyn C. Kahan¹, Vittorio Bellotti^{1†}, Philip N. Hawkins¹, Rebecca M. Myers², Martin D. Smith², Alessandra Polara², Alexander J. A. Cobb^{1†}, Steven V. Ley², J. Andrew Aquilina^{2†}, Carol V. Robinson², Isam Sharif³, Gillian A. Gray³, Caroline A. Sabin⁴, Michelle C. Jenvey⁵, Simon E. Kolstoe⁵, Darren Thompson^{5†} & Stephen P. Wood⁵

Complement-mediated inflammation exacerbates the tissue injury of ischaemic necrosis in heart attacks and strokes, the most common causes of death in developed countries. Large infarct size increases immediate morbidity and mortality and, in survivors of the acute event, larger non-functional scars adversely affect long-term prognosis. There is thus an important unmet medical need for new cardioprotective and neuroprotective treatments. We have previously shown that human C-reactive protein (CRP), the classical acute-phase protein that binds to ligands exposed in damaged tissue and then activates complement¹, increases myocardial and cerebral infarct size in rats subjected to coronary or cerebral artery ligation, respectively^{2,3}. Rat CRP does not activate rat complement, whereas human CRP activates both rat and human complement⁴. Administration of human CRP to rats is thus an excellent model for the actions of endogenous human CRP^{2,3}. Here we report the design, synthesis and efficacy of 1,6-bis(phosphocholine)-hexane as a specific small-molecule inhibitor of CRP. Five molecules of this palindromic compound are bound by two pentameric CRP molecules, crosslinking and occluding the ligand-binding B-face of CRP and blocking its functions. Administration of 1,6-bis(phosphocholine)-hexane to rats undergoing acute myocardial infarction abrogated the increase in infarct size and cardiac dysfunction produced by injection of human CRP. Therapeutic inhibition of CRP is thus a promising new approach to cardioprotection in acute myocardial infarction, and may also provide neuroprotection in stroke. Potential wider applications include other inflammatory, infective and tissue-damaging conditions characterized by increased CRP production, in which binding of CRP to exposed ligands in damaged cells may lead to complement-mediated exacerbation of tissue injury.

No inhibitors of calcium-dependent binding of human CRP to phosphocholine, the natural ligand for which it has highest affinity, were detected in a screen of the comprehensive 500,000 small-molecule library of a pharmaceutical company. We therefore rationally designed novel ligands for CRP based on the crystal structure of the CRP–phosphocholine complex⁵ and the structure of a potent drug we have developed that targets the homologous protein, serum amyloid P component (SAP)⁶. The SAP inhibitor consists of two D-proline residues linked by a hexanoyl chain, and the SAP–drug complex contains two SAP molecules crosslinked by five drug

molecules, with each D-proline located in the calcium-dependent ligand-binding pocket of an SAP protomer⁶.

In the CRP–phosphocholine complex, one phosphocholine molecule is positioned parallel to the pentamer surface on each of the five CRP subunits, and is oriented with the choline moieties, pinned between Glu 81 and Phe 66, towards the fivefold axis⁵. Two oxygen atoms of the phosphocholine phosphate group directly coordinate the two bound calcium ions of CRP, but the third is oriented away from the protein surface⁵, providing a suitable exit point to direct a crosslinking chain towards a putative twofold-axis-related subunit of an adjacent CRP pentamer. We selected a six-carbon linker to

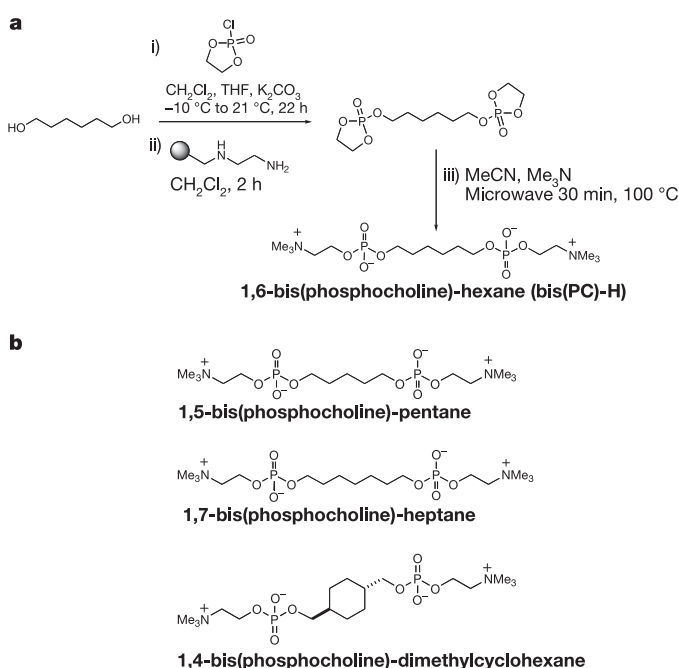


Figure 1 | Synthesis and structures of bis-phosphocholine compounds. **a**, Synthesis and structure of bis(phosphocholine)-hexane. **b**, Structures of bis(phosphocholine) compounds with different linkers. See Supplementary Information for details.

¹Centre for Amyloidosis and Acute Phase Proteins, Department of Medicine, Royal Free and University College Medical School, University College London, Rowland Hill Street, London NW3 2PF, UK. ²Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK. ³Endothelial Cell and Molecular Cardiology Group, Centre for Cardiovascular Science, University of Edinburgh, Edinburgh EH16 4TJ, UK. ⁴Department of Primary Care and Population Sciences, Royal Free and University College Medical School, University College London, London NW3 2PF, UK. ⁵School of Biological Science, University of Southampton, Southampton SO16 7PX, UK. [†]Present addresses: Addenbrookes Hospital, Cambridge CB2 2QQ, UK (G.M.H.); Dip di Biochimica, Università di Pavia, Pavia 27100, Italy (V.B.); Department of Pharmaceutical Chemistry, University of Reading, Reading RG6 6AH, UK (A.J.A.C.); School of Biological Sciences, University of Wollongong, New South Wales 2522, Australia (J.A.A.); Department of Biochemistry, University of Sussex, Falmer BN1 9QG, UK (D.T.).

connect the phosphate groups, expecting a greater separation between binding sites of a putative CRP decamer caused in part by rotation of $\sim 20^\circ$ of each CRP subunit towards the fivefold axis of the pentamer compared to SAP. This new chemical entity, composed of two phosphocholines with a hexane linker, 1,6-bis[[(trimethylammonium)ethoxy]phosphinyl]-oxy]hexane, colloquially 1,6-bis(phosphocholine)-hexane (bis(PC)-H, Fig. 1), was synthesized at high yield (see Supplementary Information).

Bis(PC)-H was bound by human CRP in a calcium-dependent manner in physiological saline solution, with a dissociation constant (K_d) of ~ 300 nM by isothermal titration calorimetry. Bis(PC)-H inhibited calcium-dependent binding of CRP to all of its other known ligands, including phosphoethanolamine, modified low-density lipoprotein, and late apoptotic or necrotic cells. The IC_{50} for inhibition by bis(PC)-H of CRP binding to immobilized pneumococcal C-polysaccharide was ~ 2 μ M, compared to ~ 20 μ M for free phosphocholine. Bis(PC)-H also blocked complement activation by human CRP and C-polysaccharide in human serum. In mixtures of drug and CRP at molar ratios between 1:1 and 1:1,000 with respect to CRP protomers, all CRP molecules became associated in pairs, as shown by size-exclusion chromatography, electron microscopy and mass spectrometry (Fig. 2). Addition of bis(PC)-H to isolated CRP, or to CRP in whole serum, resulted in disappearance of CRP immunoreactivity using a monoclonal antibody assay (Roche) specific for a calcium-dependent epitope on the binding (B) face of the CRP molecule⁷, suggesting that this epitope was occluded in the CRP–drug complex. The IC_{50} was at the ratio of three bis(PC)-H per two pentameric CRP molecules. Covalent crosslinking of CRP molecules in the CRP–drug complex, followed by fragmentation and mass spectrometric analysis (data not shown), confirmed the B face–B face association.

The X-ray crystal structure of the CRP–bis(PC)-H complex revealed two pentameric CRP molecules lying face-to-face with a common fivefold symmetry axis, crosslinked via their phosphocholine-binding sites by five drug molecules (Fig. 3; see Supplementary Information). The CRP pentamers are displaced by a relative rotation of 20° about this axis, separating the binding sites by 15 Å and inclining the axis of the bound bis(PC)-H from that of the protein complex. The structures of the protein subunits and the bound phosphocholine

component of bis(PC)-H are very similar to those observed in the CRP complex with free phosphocholine⁵ (r.m.s. best fit over all C α = 0.4 Å). However, a considerably worse fit (r.m.s. all C α = 0.8 Å) was observed upon overlaying complete pentamers. This was due to a systematic movement of β -strands by ~ 1 Å towards the fivefold axis, resulting in a contraction of ~ 2 Å in the pore diameter at the centre of the pentamer in the CRP–bis(PC)-H complex. There are $\sim 25\%$ more inter-atomic contacts of less than 3.5 Å per CRP subunit than in the complex with phosphocholine. Identical cryopreservation methods were used during data collection, indicating that this effect is a specific consequence of bis(PC)-H binding by CRP. There are also two additional calcium ions bound per subunit: one by the side chain of Asp 60 and the carbonyl of Asn 59, close to the ligand-binding double calcium site, and the other by the main-chain carbonyl of Glu 70 and the side-chain carboxylate of the same residue from the twofold-symmetry-related subunit of the adjacent pentamer, providing decamer stabilization. The ligand fields of these calcium ions are limited, and they are likely to be occupied only by virtue of the high calcium ion concentration (50 mM) in the crystallization cocktail. Further decamer stability is provided by ten inter-pentamer ion pairs between Lys 69 N ϵ and Glu 85' O δ . All of these interactions between pentamers are orthogonal to the observed direction of contraction, and are unlikely to be its cause.

The electron density for the phosphocholine component of bis(PC)-H and the first carbon of the crosslinker is very good but, not unexpectedly for such a flexible linker, the density for the central four carbon atoms is weaker (Fig. 3c). In order to fit these atoms within the space available and to achieve the required approach path to the phosphate groups, there is an unfavourable eclipsed rotamer about the C3–C4 bond of the linker. However, 1,5-bis(phosphocholine)-pentane, with a shorter linker, was bound with substantially lower affinity ($K_d \sim 3.7$ μ M), and there was no binding at all to the more rigid bis(phosphocholine)-dimethylcyclohexane (Fig. 1 and Supplementary Information). In contrast 1,7-bis(phosphocholine)-heptane (Fig. 1; $K_d \sim 300$ nM) evidently had sufficient linker length and flexibility to permit this mechanism of drug–protein interaction.

Bolus intravenous or intraperitoneal injections in mice and rats of up to 1 mmol kg⁻¹ bis(PC)-H in physiological saline solution, or continuous infusion at 1 mmol kg⁻¹ per day for seven days via

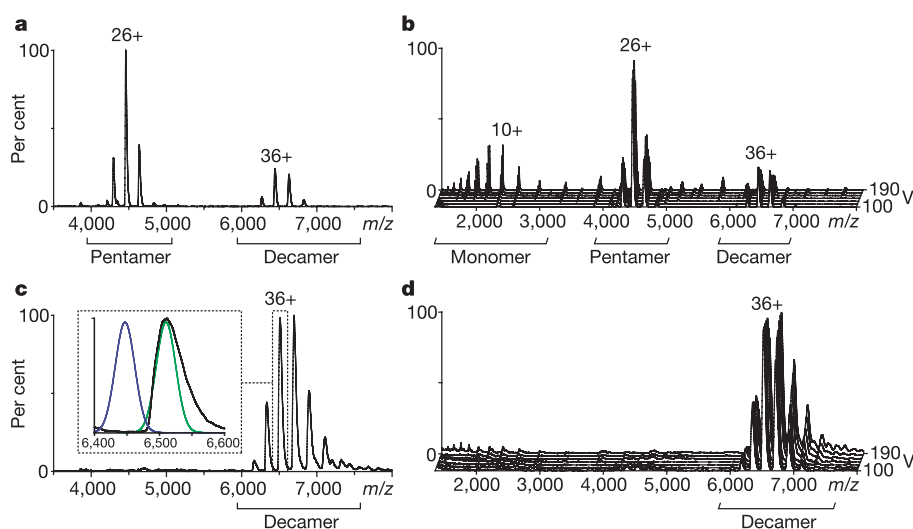


Figure 2 | Nano electrospray mass spectra of CRP. **a, b**, CRP alone yielded peaks corresponding to the native pentamer and traces of decamer, reflecting the known tendency of native CRP to aggregate. Accelerating the ions (with increasing cone voltage in the atmospheric pressure region of the mass spectrometer) increases their internal energy and causes the decamers and pentamers to dissociate, yielding predominantly free protomers at the highest cone voltages. **c, d**, Addition of bis(PC)-H yields exclusively

decameric CRP, which is resistant to gas phase collision-induced dissociation. The peaks in the mass spectrum are shifted towards higher m/z values, indicating drug binding. The inset in **c** compares the 36+ charge state in the mass spectrum of the decamer in the absence of bis(PC)-H (blue) and the peak anticipated for the addition of five bis(PC)-H molecules (green), demonstrating the stoichiometry of binding (five bis(PC)-H molecules and two pentameric CRP molecules).

subcutaneous osmotic pump, were tolerated without noticeable adverse effects, and inhibited binding of injected human CRP to other ligands and its reactivity using the Roche assay⁷. The plasma half-life of bis(PC)-H in mice was ~ 90 min. Continuous infusion of 1 mmol bis(PC)-H per kg per day in rats completely blocked the effects of daily subcutaneous injections of 40 mg kg⁻¹ human CRP (~ 1.74 μ mol CRP protomer), despite the presence of rat CRP, which binds the drug with $K_d \sim 150$ nM, circulates at 300–500 mg l⁻¹, and is produced at the rate of ~ 10 μ mol protomer per kg per day.

Clinical treatment with a CRP inhibitor could be started immedi-

ately upon admission to hospital following acute myocardial infarction—this would precede the acute phase CRP response, which starts about 6 h after onset of pain and peaks at about 50 h (refs 8, 9). We therefore initiated infusion of bis(PC)-H before coronary artery ligation in rats, and gave the first of five daily subcutaneous injections of human CRP immediately after recovery from surgery, closely replicating the initial dynamics of the endogenous human CRP response. Administration of human CRP was associated with increased mortality compared to vehicle-only controls, as we have previously reported². In contrast, there were no deaths among the rats receiving bis(PC)-H in addition to CRP (Table 1) (Fisher's exact

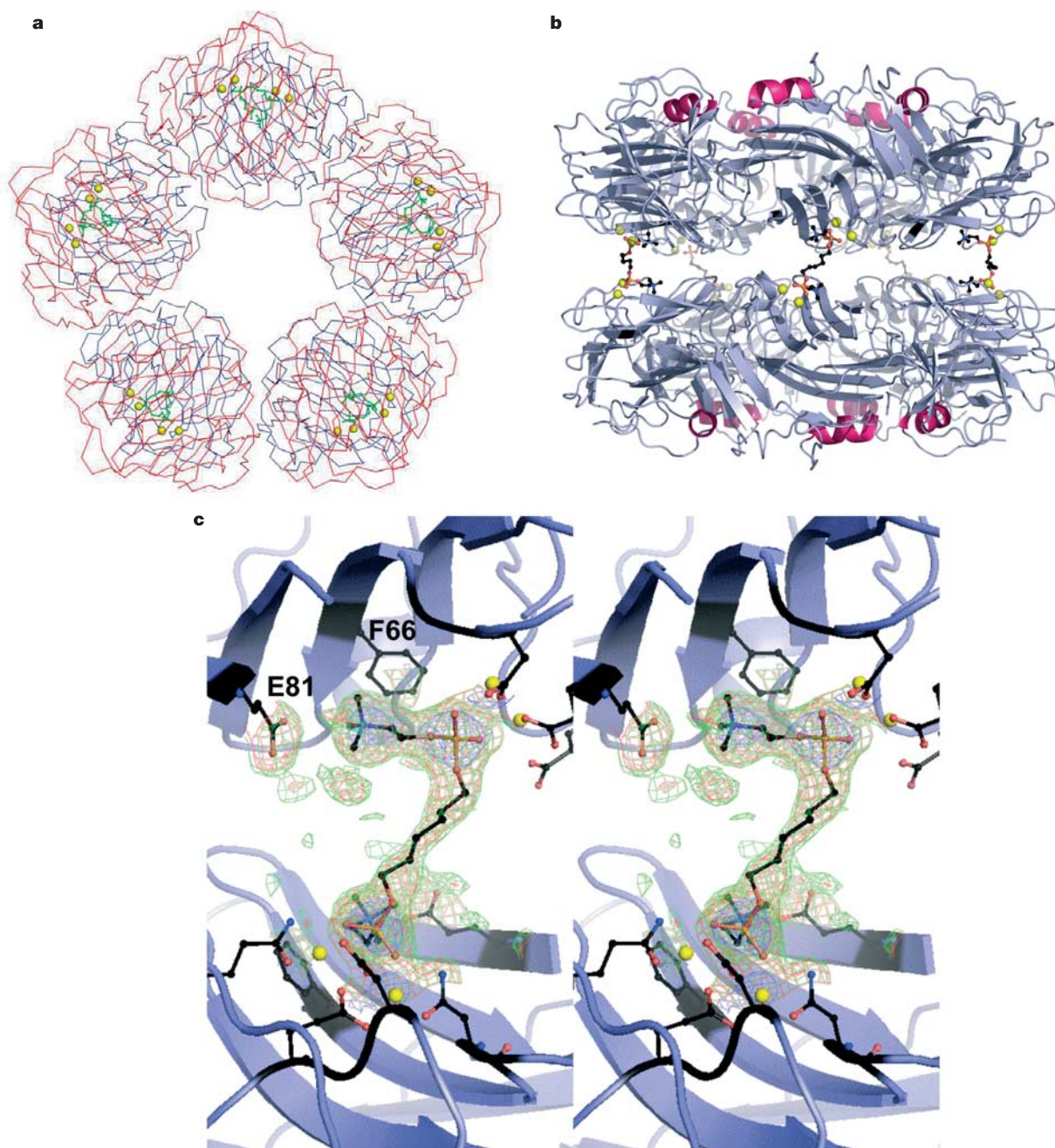


Figure 3 | Structure of the complex of CRP with 1,6-bis(phosphocholine)-hexane. **a**, Two CRP pentamers viewed down the fivefold axis, one in red and one in blue, crosslinked via their B-faces by five molecules of bis(PC)-H (green). **b**, View of the complex perpendicular to the fivefold axis and along the local twofold axis, relating pairs of subunits. A-face helices are in pink. Both views show the rotation of pentamers relative to each other, with corresponding displacement of calcium ions (yellow) and inclination of the

drug molecules. **c**, Stereo view of electron density ($2|F_o| - |F_c|$) contoured at 3σ (blue), 1σ (red) and 0.75σ (green), and the fitted bis(PC)-H molecule showing how the buckled linker chain directs the phosphocholine head group into the double calcium site, with the choline moiety sandwiched between Phe 66 and Glu 81, and two phosphate oxygens coordinating the calcium ions. Images were prepared using PyMol²³.

Table 1 | Effect of human CRP and bis(PC)-H on of myocardial infarct size and cardiac dysfunction

Group (n)	Sham-operated (5)	Coronary artery ligation			P values		
		A (12) Vehicle	B (15) CRP	C (11) CRP+Bis(PC)-H	A versus B	A versus C	B versus C
Treatment	None						
Infarct size (p) (%)	0	17.0 ± 3.8	24.8 ± 3.3	18.6 ± 3.8	0.0001	0.32	0.0002
Infarct size (w) (%)	0	17.7 ± 4.7	29.5 ± 5.7	18.4 ± 3.7	0.0001	0.68	0.0001
ST voltage (μV)	-44.8 ± 24.1	14.9 ± 82.3	95.0 ± 61.8*	16.2 ± 74.3	0.006	0.97	0.008
Elevated ST, n (%)	0	7 (58.3)	14 (100)*	7 (63.6)	0.007	1.00	0.06
Ejection fraction (%)	72.67 ± 3.91	37.41 ± 15.95	28.75 ± 14.82	38.40 ± 11.47	0.16	0.87	0.08
LVEDP (mm Hg)	4.4 ± 1.8	8.8 ± 4.1	12.9 ± 3.8	4.2 ± 6.3	0.01	0.05	0.0002
LVEDD (mm)	6.54 ± 0.24	7.43 ± 0.87	8.06 ± 0.74	7.19 ± 1.08	0.05	0.56	0.02
LVEDS (mm)	3.1 ± 0.51	5.94 ± 1.39	6.70 ± 1.21	5.60 ± 1.16	0.14	0.53	0.03

Values shown are mean ± s.d. except for ST elevation, where the number of individuals with ST elevation and their proportion as a percentage of each group are shown. P values are from unpaired *t*-tests. Infarct size on day 5 is shown as a percentage of the left ventricle, measured by planimetry (p) and by weight (w). VEDD, left ventricular end diastolic diameter; LVEDP, left ventricular end diastolic pressure; LVEDS, left ventricular end systolic diameter; ST, ST segment of the electrocardiogram.

* *n* = 14 for these measurements.

test for comparison of mortality in all groups, *P* = 0.08). Infarct size on day 5 was substantially larger in the rats treated with CRP (unpaired *t*-test, *P* = 0.0001 compared to vehicle-treated rats), but in rats receiving bis(PC)-H as well as CRP, infarct size was the same as in vehicle-only controls (Table 1). Electro- and echocardiographic indices of cardiac function on day 5 were consistent with the larger infarcts in CRP-treated rats and with a protective effect of bis(PC)-H (Table 1).

At bleed-out on day 5, 24 h after the last dose of human CRP, the mean (±s.d.) concentration of human CRP in the serum of CRP-treated rats was $16.7 \pm 10.6 \text{ mg l}^{-1}$, but human CRP was detectable by standard immunoassay in only 4 of the 11 CRP-treated rats receiving bis(PC)-H (mean $3.3 \pm 1.5 \text{ mg l}^{-1}$), and was not detected at all by the Roche assay, demonstrating complex formation between bis(PC)-H and CRP. Continuous infusion of bis(PC)-H thus resulted in accelerated clearance of human CRP and blocked its function.

In a separate experiment, administration of the same dose of bis(PC)-H to rats (*n* = 11) that underwent coronary artery ligation but did not receive human CRP had no effect on infarct size at day 5, compared to coronary artery ligation controls receiving vehicle alone (*n* = 6). The mean infarct size (±s.d.) as a percentage of the left ventricle was $19.0 \pm 1.9\%$ for the bis(PC)-H group and $20.0 \pm 2.0\%$ for the vehicle group. There was also no difference in haemodynamic measures, indicating that bis(PC)-H had no significant cardiovascular effect in the absence of human CRP.

The physiological role of human CRP is unknown because no deficiency or structural polymorphism in human CRP, or experimental CRP knockout, has yet been reported. Experimental animal studies suggest that CRP may contribute to innate immunity, can be anti-inflammatory, and may protect against autoimmunity, and we have shown that the administration of human CRP exacerbates pre-existing tissue damage in a complement-dependent fashion^{2,3}. Previously reported pro-inflammatory effects of human CRP preparations on cells *in vitro* were due to bacterial endotoxin and other contaminants rather than CRP itself^{10–12}, and pure human CRP is not pro-inflammatory when injected into healthy animals^{12,13}. Furthermore, transgenic human CRP is not pro-atherogenic or pro-thrombotic in apolipoprotein-E-knockout mice¹⁴. However, the evolutionary conservation of CRP does not exclude potentially harmful effects—although CRP may have evolved to promote beneficial functions, it might also enhance lesion severity, especially in post-reproductive-age diseases such as atherothrombosis, auto-immune and other chronic inflammatory conditions, which are not subject to evolutionary pressure. Our rationally designed CRP-targeting drug demonstrates that CRP inhibition is a valid therapeutic strategy that is unlikely to have adverse effects, and that may prove informative about the physiological and pathological roles of human CRP.

METHODS

Reagents and assays. Human CRP was isolated from malignant ascites fluid as reported^{3,13}. Human CRP was assayed by the Roche⁷ and Dade-Behring methods¹⁵ and by electroimmunoassay¹⁶. Rat CRP and complement C3 were measured by electroimmunoassay⁴. Calcium-dependent binding of ¹²⁵I-labelled CRP to pneumococcal C-polysaccharide (Statens Serum Institut) and modified human low-density lipoprotein¹⁷, which were covalently immobilized on Corning Costar *N*-hydroxysuccinimide microtitre plates, was compared in the presence and absence of inhibitor compounds. Binding of CRP to phosphoethanolamine-Sepharose was determined as reported previously for SAP¹⁸. Activation of complement in whole human serum by CRP and C-polysaccharide in the presence and absence of bis(PC)-H was monitored by two-dimensional immunoelectrophoresis⁴ with monospecific antiserum against human C3.

Binding affinity of CRP for ligands in solution in 0.01 M Tris, 0.14 M NaCl, 0.002 M CaCl₂, 0.1% NaN₃ pH 8.0 (TC buffer plus azide) was measured at 37 °C by isothermal titration calorimetry⁶. The effects of bis(PC)-H on CRP molecules were monitored by chromatography on a Superdex 200 HR10/30 column in the ÄKTA Explorer 100 HPLC system (Amersham Biosciences) eluted with TC buffer, by uranyl acetate negative-staining electron microscopy on carbon grids, and by electrospray mass spectrometry (Supplementary Information).

Myocardial infarction. ALZET osmotic mini-pumps, delivering $10 \mu\text{l h}^{-1}$ for seven days, were implanted subcutaneously in male Wistar rats (200–220 g) two days before coronary artery ligation. Groups A and B received TC buffer; and group C received 1.0 M bis(PC)-H in TC buffer, providing 1 mmol kg^{-1} per day. Coronary artery ligation or sham operations were performed under intra-peritoneal anaesthesia with 75 mg kg^{-1} ketamine, 0.6 mg kg^{-1} xylazine and 0.2 mg kg^{-1} atropine^{19,20}, and post-operative atipamezole 0.5 mg kg^{-1} two days after pump implantation. Five daily subcutaneous injections of either TC buffer alone (group A) or human CRP at 40 mg kg^{-1} per day in TC buffer (groups B and C) were given, starting immediately after recovery from coronary surgery. Echocardiography (10–22 MHz probe, Dynamic Imaging)^{21,22}, right carotid artery cannulation (using a pressure-transducer tipped catheter 1.4 F, Millar Instrument Inc.) and cardiac catheterization were performed on day 5, with the rats under isofluorane anaesthesia. The rats were then bled, and their hearts excised, cleaned, weighed and frozen. Frozen hearts were then cut transversely into 2.5-mm slices and stained with 1% (w/v) 2,3,5-triphenyl tetrazolium chloride in phosphate buffer. Infarct size was measured by planimetry (using an MCID image analysis system, Imaging Research Inc.) on formalin-fixed, glass-mounted sections, and confirmed by dissection and weighing. All treatments and measurements were performed by an experimenter blind to the treatment group.

Received 22 September 2005; accepted 21 February 2006.

1. Pepys, M. B. & Hirschfield, G. M. C-reactive protein: a critical update. *J. Clin. Invest.* **111**, 1805–1812 (2003).
2. Griselli, M. et al. C-reactive protein and complement are important mediators of tissue damage in acute myocardial infarction. *J. Exp. Med.* **190**, 1733–1739 (1999).
3. Gill, R., Kemp, J. A., Sabin, C. & Pepys, M. B. Human C-reactive protein increases cerebral infarct size after middle cerebral artery occlusion in adult rats. *J. Cereb. Blood Flow Metab.* **24**, 1214–1218 (2004).
4. de Beer, F. C. et al. Isolation and characterisation of C-reactive protein and serum amyloid P component in the rat. *Immunology* **45**, 55–70 (1982).
5. Thompson, D., Pepys, M. B. & Wood, S. P. The physiological structure of human C-reactive protein and its complex with phosphocholine. *Structure* **7**, 169–177 (1999).

6. Pepys, M. B. *et al.* Targeted pharmacological depletion of serum amyloid P component for treatment of human amyloidosis. *Nature* **417**, 254–259 (2002).
7. Eda, S., Kaufmann, J., Roos, W. & Pohl, S. Development of a new microparticle-enhanced turbidometric assay for C-reactive protein with superior features in analytical sensitivity and dynamic range. *J. Clin. Lab. Anal.* **12**, 137–144 (1998).
8. Kushner, I., Broder, M. L. & Karp, D. Control of the acute phase response. Serum C-reactive protein kinetics after acute myocardial infarction. *J. Clin. Invest.* **61**, 235–242 (1978).
9. de Beer, F. C. *et al.* Measurement of serum C-reactive protein concentration in myocardial ischaemia and infarction. *Br. Heart J.* **47**, 239–243 (1982).
10. Taylor, K. E., Giddings, J. C. & van den Berg, C. W. C-reactive protein-induced *in vitro* endothelial cell activation is an artefact caused by azide and lipopolysaccharide. *Arterioscler. Thromb. Vasc. Biol.* **25**, 1225–1230 (2005).
11. Pepys, M. B. CRP or not CRP? That is the question. *Arterioscler. Thromb. Vasc. Biol.* **25**, 1091–1094 (2005).
12. Pepys, M. B. *et al.* Pro-inflammatory effects of bacterial recombinant human C-reactive protein are caused by contamination with bacterial products not by C-reactive protein itself. *Circ. Res.* **97**, e97–103 (2005).
13. Clapp, B. R. *et al.* Inflammation and endothelial function: direct vascular effects of human C-reactive protein on nitric oxide bioavailability. *Circulation* **111**, 1530–1536 (2005).
14. Hirschfield, G. M. *et al.* Transgenic human C-reactive protein is not proatherogenic in apolipoprotein E-deficient mice. *Proc. Natl Acad. Sci. USA* **102**, 8309–8314 (2005).
15. Ledue, T. B. *et al.* Analytical evaluation of particle-enhanced immunonephelometric assays for C-reactive protein, serum amyloid A and mannose-binding protein in human serum. *Ann. Clin. Biochem.* **35**, 745–753 (1998).
16. Nelson, S. R. *et al.* Serum amyloid P component in chronic renal failure and dialysis. *Clin. Chim. Acta* **200**, 191–200 (1991).
17. Bhakdi, S., Torzewski, M., Klouche, M. & Hemmes, M. Complement and atherogenesis. Binding of CRP to degraded, nonoxidized LDL enhances complement activation. *Arterioscler. Thromb. Vasc. Biol.* **19**, 2348–2354 (1999).
18. Hawkins, P. N., Tennent, G. A., Woo, P. & Pepys, M. B. Studies *in vivo* and *in vitro* of serum amyloid P component in normals and in a patient with AA amyloidosis. *Clin. Exp. Immunol.* **84**, 308–316 (1991).
19. Gray, G. A., Mickley, E. J., Webb, D. J. & McEwan, P. E. Localization and function of ET-1 and ET receptors in small arteries post-myocardial infarction: upregulation of smooth muscle ET_B receptors that modulate contraction. *Br. J. Pharmacol.* **130**, 1735–1744 (2000).
20. Miller, A. A., Megson, I. L. & Gray, G. A. Inducible nitric oxide synthase-derived superoxide contributes to hyperactivity in small mesenteric arteries from a rat model of chronic heart failure. *Br. J. Pharmacol.* **131**, 29–36 (2000).
21. Mora, A. *et al.* Deficiency of PDK1 in cardiac muscle results in heart failure and increased sensitivity to hypoxia. *EMBO J.* **22**, 4666–4676 (2003).
22. Denvir, M. A. *et al.* Influence of scanning frequency and ultrasonic contrast agent on reproducibility of left ventricular measurements in the mouse. *J. Am. Soc. Echocardiogr.* **18**, 155–162 (2005).
23. DeLano, W. L. The PyMOL Molecular Graphics System <http://www.pymol.org> (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Ganellin for a preliminary synthesis of bis(PC)-H by a different route, D. Moore, M. Neal, M. Singer and R. Stidwell for assistance with rat studies, and P. Vallance for advice. The study was funded by an MRC Programme Grant (M.B.P. and P.N.H.), a project grant from the US NIH National Heart, Lung and Blood Institute's Inflammation and Thrombosis Program (M.B.P.), The Wellcome Trust (S.P.W.) and Pentraxin Therapeutics Ltd. G.M.H. held an MRC Clinical Research Training Fellowship, M.D.S. holds a Royal Society University Research Fellowship and S.V.L. a Novartis Research Fellowship.

Author Contributions The study was initiated, designed and directed by M.B.P., and the paper was written and the novel compounds were designed by M.B.P. and S.P.W. *In vitro* and *in vivo* experimental work was performed M.B.P., G.M.H., G.A.T., J.R.G., M.C.K. and V.B. P.N.H. contributed to study design. Chemical syntheses were designed and performed by S.V.L., R.M.M., M.D.S., A.P. and A.J.A.C., and mass spectrometry was performed by J.A.A. and C.V.R. Myocardial infarction experiments were conducted by G.A.G. and I.S., C.A.S. performed statistical analyses, and S.P.W., M.C.J., S.E.K. and D.T. did the X-ray crystallography.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.B.P. (m.pepys@medsch.ucl.ac.uk).

LETTERS

Lysyl oxidase is essential for hypoxia-induced metastasis

Janine T. Erler¹, Kevin L. Bennewith¹, Monica Nicolau², Nadja Dornhöfer⁴, Christina Kong³, Quynh-Thu Le¹, Jen-Tsan Ashley Chi⁵, Stefanie S. Jeffrey² & Amato J. Giaccia¹

Metastasis is a multistep process responsible for most cancer deaths, and it can be influenced by both the immediate micro-environment (cell–cell or cell–matrix interactions) and the extended tumour microenvironment (for example vascularization)¹. Hypoxia (low oxygen) is clinically associated with metastasis and poor patient outcome, although the underlying processes remain unclear². Microarray studies have shown the expression of lysyl oxidase (LOX) to be elevated in hypoxic human tumour cells³. Paradoxically, LOX expression is associated with both tumour suppression and tumour progression, and its role in tumorigenesis seems dependent on cellular location, cell type and transformation status^{4–9}. Here we show that LOX expression is regulated by hypoxia-inducible factor (HIF) and is associated with hypoxia in human breast and head and neck tumours. Patients with high LOX-expressing tumours have poor distant metastasis-free and overall survivals. Inhibition of LOX eliminates metastasis in mice with orthotopically grown breast cancer tumours. Mechanistically, secreted LOX is responsible for the invasive properties of hypoxic human cancer cells through focal adhesion kinase activity and cell to matrix adhesion. Furthermore, LOX may be required to create a niche permissive for metastatic growth. Our findings indicate that LOX is essential for hypoxia-induced metastasis and is a good therapeutic target for preventing and treating metastases.

First we validated LOX as a hypoxia-responsive gene and found it to be regulated at the messenger RNA level by hypoxia-inducible factor-1 (HIF-1) through a functional hypoxia-responsive element identified and tested in the LOX promoter (Supplementary Fig. 1). To examine the clinical relevance of hypoxia-induced LOX, we performed a retrospective breast cancer and a prospective head and neck cancer study^{10,11}. There was significant correlation between the LOX expression level and hypoxia in patients (Fig. 1a, d). LOX expression was statistically associated with oestrogen receptor (ER) status (Supplementary Table 1), consistent with the finding that highly hypoxic tumours are most likely to be ER-negative¹². This is of clinical importance because ER-negative breast cancer patients generally have a worse prognosis¹³. LOX expression was associated with lower distant metastasis-free survival and overall survival in breast cancer patients with ER-negative tumours, and in head and neck cancer patients (Fig. 1b–f).

To study the way in which hypoxia-induced LOX could influence metastasis and survival, we generated LOX short hairpin RNA (shRNA) expressing human MDA231 breast and SiHa cervical cancer cells. These cells expressed significantly less LOX mRNA and protein than cells expressing a scrambled control sequence, but grew at similar rates *in vitro* and *in vivo* (Supplementary Fig. 4). MDA231 cells were grown as orthotopic tumours in nude mice. Hypoxic

regulation of LOX was confirmed in tumours by staining for LOX and pimonidazole (Fig. 2a). Mice bearing shRNA-expressing tumours had significantly fewer lung metastases and no liver metastases, in contrast with wild-type tumours (Fig. 2b, c).

To evaluate the therapeutic usefulness of inhibiting LOX, mice

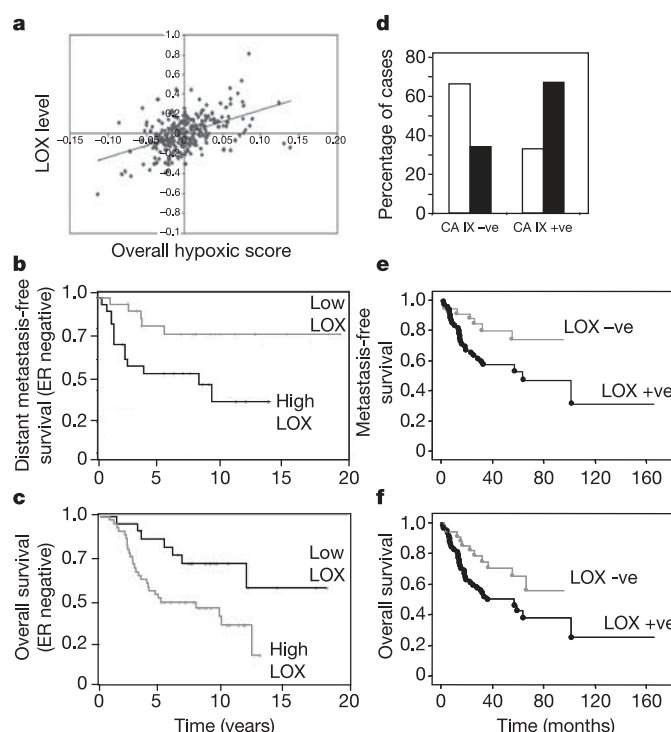


Figure 1 | Cancer patients expressing high levels of LOX have poor outcome. **a**, Correlation between the LOX expression level (*y* axis) and average hypoxia score (*x* axis) among breast cancer patients ($n = 295$)¹⁰. $P < 0.0001$. **b**, **c**, Kaplan–Meyer plots showing that patients with ER-negative breast tumours with high LOX expression levels had statistically significant reduced metastasis-free survival (**b**; $P = 0.009$) and overall survival (**c**; $P = 0.015$) than patients with low LOX expression levels. **d**, Comparison of LOX protein expression levels with those of CA-IX in a tissue array study from head and neck cancer patients ($n = 91$)¹¹. Filled bars, LOX positive; open bars, LOX negative. $P = 0.006$. **e**, **f**, Kaplan–Meyer plots showing that head and neck cancer patients whose tumours stained positive for LOX had statistically significant reduced metastasis-free survival (**e**; $P = 0.02$) and overall survival (**f**; $P = 0.046$) than patients whose tumours stained negative for LOX.

¹Department of Radiation Oncology, ²Department of Surgery, ³Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA. ⁴Department of Obstetrics and Gynecology, University of Leipzig, Leipzig 04103, Germany. ⁵Department of Molecular Genetics and Microbiology, Duke University, Durham, North Carolina 27708, USA.

with control MDA231 tumours were treated with β -aminopropionitrile (BAPN), a specific and irreversible inhibitor of LOX enzymatic activity. Because BAPN is reported to have effects on other LOX family members (of which there are four, with some overlap in structure and function¹⁴), we also treated mice with our purified LOX antibody, which cross-reacts with murine LOX, allowing us to assess deleterious side effects. Mice treated with BAPN or 20 mg kg⁻¹ antibody did not have any lung metastases or liver metastases (Fig. 2b, c). Mice that received reduced antibody doses or periods of BAPN treatment displayed significantly fewer lung metastases and no liver metastases, even when metastases had already formed (Fig. 2d). Inhibition of LOX (with shRNA, BAPN or antibody) did not have a major effect on primary tumour growth, and there was no association between tumour size and the number of metastases (data not shown). We observed reduced LOX activity levels in the blood of BAPN-treated and antibody-treated mice, indicating that the

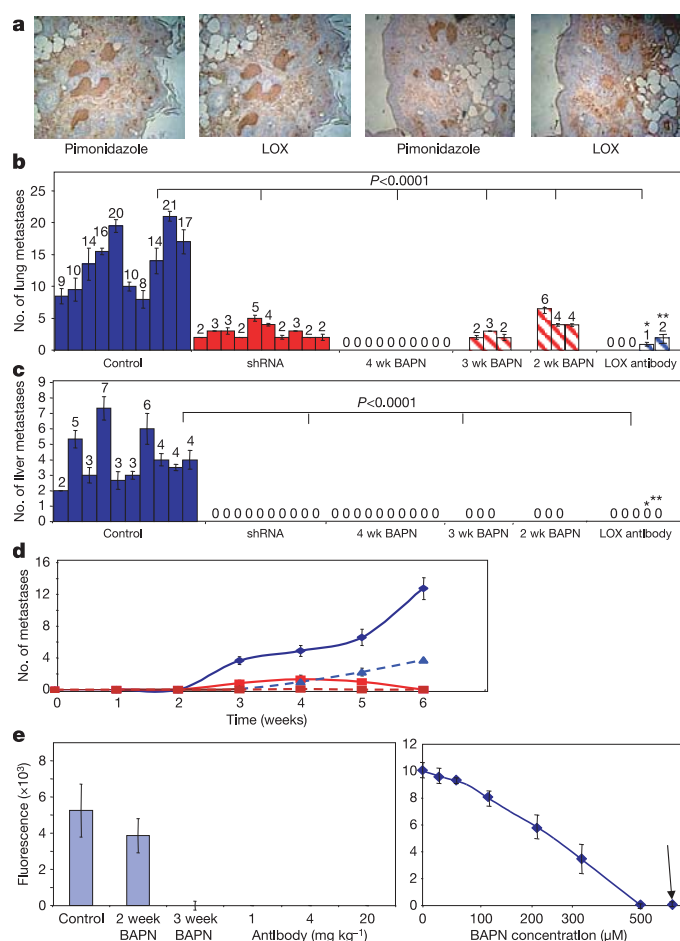


Figure 2 | Inhibition of LOX prevents metastasis *in vivo*. **a**, Two pairs of serial sections of orthotopic MDA231 tumours stained for pimonidazole or LOX. Positive staining is brown (haematoxylin/eosin stain, blue background). Original magnification, $\times 10$. **b**, **c**, Microscopic quantification of metastases in lungs and livers stained with hematoxylin and eosin. Data are numbers of metastases formed at the end of the six-week experiment per mouse (means $\pm$ s.e.m.) for the ten step sections, based on three independent experimental repeats. Antibody doses: unlabelled, 20 mg kg⁻¹; asterisk, 4 mg kg⁻¹; two asterisks, 1 mg kg⁻¹. Control, $n = 10$; LOX shRNA, $n = 10$; 4 wk BAPN, $n = 10$; 3 wk BAPN, $n = 3$; 2 wk BAPN, $n = 3$; LOX antibody, $n = 5$. Treatment details are given in Methods. **d**, Metastases counted weekly over six weeks for ten sections ($n = 5$ mice). Dark blue solid line, control lung; blue dashed line, control liver; red solid line, BAPN lung; dark red dashed line, BAPN liver. **e**, LOX fluorescent activity assays: left, *in vivo* enzymatic assay; right, *in vitro* enzymatic assay (arrowed point indicates 20 mg kg⁻¹ LOX antibody). All data are plotted as means $\pm$ s.e.m.

antibody might prevent LOX enzymatic activity, which was confirmed in an *in vitro* assay (Fig. 2e). These results provide powerful evidence that targeting a hypoxia-related protein can prevent metastasis in a breast cancer model.

The first step of the metastatic process is cell invasion. Primary tumours from control, BAPN-treated or antibody-treated mice showed evidence of invasion, whereas tumours expressing LOX shRNA did not (Fig. 3a). We therefore examined the invasion of various human cancer cells *in vitro*. All cell lines investigated showed significantly increased invasion under hypoxia or anoxia compared with normoxic cells (Fig. 3b, and Supplementary Fig. 5b–e). This could be prevented by treatment with LOX antisense oligonucleotides (which decreased LOX mRNA expression; Supplementary Fig. 5a), BAPN, LOX antibody or shRNA expression, but not with LOX sense oligonucleotides (Fig. 3b, and Supplementary Fig. 5b–e). Reduction of *in vitro* invasion in air by LOX antisense or BAPN treatment is consistent with a previous report⁸. Invasion could be restored in shRNA cells through the overexpression of mature human LOX (which overcame shRNA inhibition; Supplementary Fig. 4a) but not with transfection of a catalytically inactive LOX gene (delta cat). Invasion was also increased by the addition of conditioned medium (CM) from control cells or shRNA cells transfected with

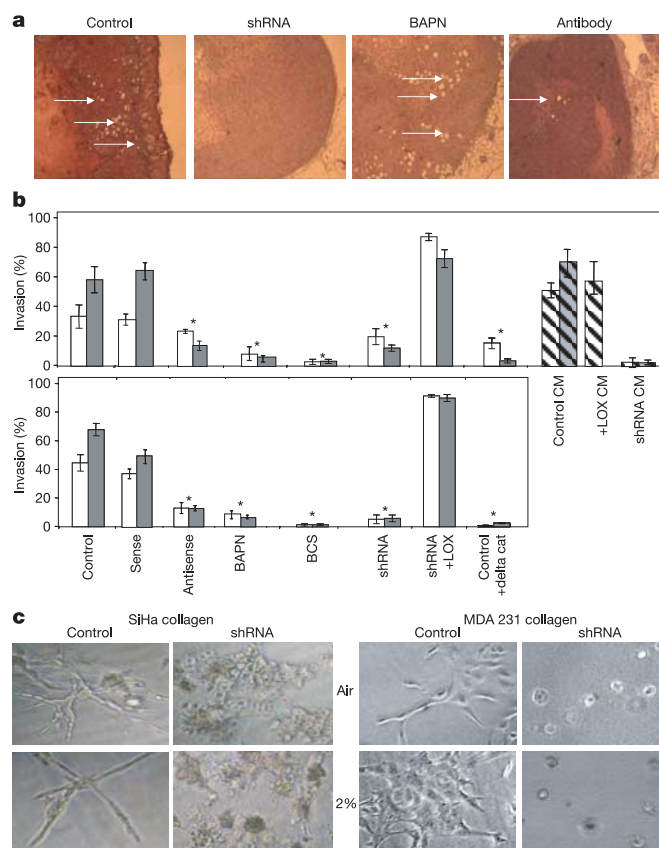


Figure 3 | Inhibition of LOX prevents hypoxia-stimulated increased cell invasion. **a**, Haematoxylin/eosin staining from the orthotopic tumour study (control, shRNA, 4 wk BAPN and 20 mg kg⁻¹ antibody). Evidence of invasion is visible through the presence of fat cells within the tumour mass (white arrows). Original magnification, $\times 4$. **b**, *In vitro* invasion of MDA231 (top) and SiHa (bottom) cells (see Methods for details). Results are representative of three independent experimental repeats (means $\pm$ s.e.m.). Asterisks indicate significant difference ($P < 0.01$) from control cells (Student's *t*-test). White background, air; grey background, hypoxia. **c**, Control or LOX shRNA-expressing cells were grown in type I collagen. Note the invasive morphology (EMT-like branching) in control cells and the more rounded, spheroid-like morphology in LOX shRNA-expressing cells. Similar results were obtained with collagen IV. Original magnification, $\times 30$.

LOX, but not by the addition of CM from aerobic or hypoxic shRNA cells (Fig. 3b). Because LOX is a copper-dependent enzyme that can act intracellularly or extracellularly, cells were treated with bathocuprione disulphonate (BCS), a copper chelator that cannot permeate the cell membrane¹⁵. This inhibition of extracellular LOX prevented the enhanced invasion of hypoxic cells (Fig. 3b). Neither treatment with BAPN nor treatment with BCS affected cell viability (data not shown), and we observed no differences in histone acetylation in our cell lines with the different treatments, ruling out intracellular LOX involvement (data not shown). Taken together, these results demonstrate a role for enzymatically active secreted LOX in the enhanced *in vitro* invasion observed in oxygen-deprived cells.

The ability of cells to reorganize and contract three-dimensional type I collagen gels is regarded as an *in vitro* model for invasion¹⁶. In air, control cells showed a branching morphogenesis typical of invasive human cancer cells grown on collagen, which was enhanced by hypoxia (Fig. 3c). The LOX shRNA cells grew in a markedly different manner, remaining in spheroid-like cell clusters, showing little branching in air or hypoxia. These results strongly indicate a role for LOX in the development of an invasive phenotype of hypoxic human cancer cells *in vitro*.

Acquisition of motility is required before cells can migrate and invade. Invasive cell migration is a multi-step process¹⁷ that commences with pseudopod protrusion at the leading edge driven by actin polymerization, resulting in focal adhesion formation and the activation of integrin and focal adhesion kinase (FAK) (Supplementary Fig. 6a). We observed intense immunofluorescent staining of extracellular LOX at the leading edge of MDA231 cells grown on collagen, particularly in hypoxic conditions (Fig. 4a). LOX protein expression extended along hairlike fibres protruding from the cell surface into the collagen matrix. This was not observed in the LOX shRNA cells, which showed levels of intracellular LOX expression equivalent to the controls containing no antibody (data not shown). We observed remodelling of the actin cytoskeleton with increased formation of stress fibres and focal adhesion in MDA231 control cells particularly in hypoxia, which were not seen in cells expressing LOX shRNA (Fig. 4b). Hypoxia additionally increased FAK phosphorylation (activation) in control cells (consistent with previous reports) but not in the LOX shRNA cells (Fig. 4c).

We sought a role for fibronectin (FN), because it was strongly associated with LOX expression in the breast cancer microarray data set we analysed (Supplementary Figs 2 and 3), is reported to regulate invasion, bind to integrins and interact with LOX, and is hypoxia-induced^{18,19}. However, neither FAK phosphorylation nor cell invasion was affected by transfection with FN antisense oligonucleotides or by the addition of plasma FN (pFN) or cellular FN (cFN) (which do and do not interact with LOX, respectively¹⁹; Supplementary Fig. 6b, and data not shown). Moreover, FAK phosphorylation was intact in FN-null cells but was decreased by transfection with LOX antisense oligonucleotides, which strongly inhibited invasion (similarly to BAPN addition; Supplementary Fig. 6c). FAK phosphorylation could be induced in the cells expressing LOX shRNA by transfection with mature LOX (but not catalytically dead LOX), confirming a role for LOX in FAK phosphorylation (Fig. 4c). This could be prevented by the addition of BAPN or catalase (consistent with a recent report⁹), or a blocking antibody against β_1 integrin but not against α_6 integrin (Fig. 4c, and Supplementary Fig. 6b). These results show an enzymatic role for LOX in the regulation of FAK through β_1 integrin, in a FN-independent manner. We propose that this is because LOX increases fibrillar collagen, which is a ligand for β_1 integrin. Other integrin pathways have been reported to mediate hypoxia-induced invasion²⁰. Complete elimination of FAK activity in hypoxic control cells was achieved when both β_1 integrin activation and hydrogen peroxide production (a by-product of LOX activity) were prevented (Supplementary Fig. 6b), demonstrating the importance of both these mechanisms in hypoxia.

Increased adhesion is a characteristic of invasive cells with a

mesenchymal phenotype and is essential for their motility. Both the MDA231 and SiHa cells expressing LOX shRNA showed decreased adhesion to collagen I (Fig. 4d, and Supplementary Fig. 6d) and Matrigel (data not shown), which could be restored on transfection with mature LOX. In invasive migration, increased adhesion additionally results in the recruitment of proteases (such as matrix metalloproteinases (MMPs)) to the attachment sites. Although expression levels of MMP-2 and MMP-9 were elevated by hypoxia, consistent with previous reports²¹, and MMP-2 and MMP-14 expression was strongly correlated with LOX in breast cancer patients (Supplementary Figs 2 and 3 and Supplementary Tables 2 and 3), LOX expression did not affect MMP activities (data not shown). However, time-lapse photography of wild-type cells and cells expressing LOX shRNA within a collagen matrix or subjected to scratch assays revealed that LOX inhibition completely prevented cell movement (Fig. 4e, f). Taken together, these results demonstrate a crucial role for LOX in cell motility through its effects on cell adhesion through FAK activation.

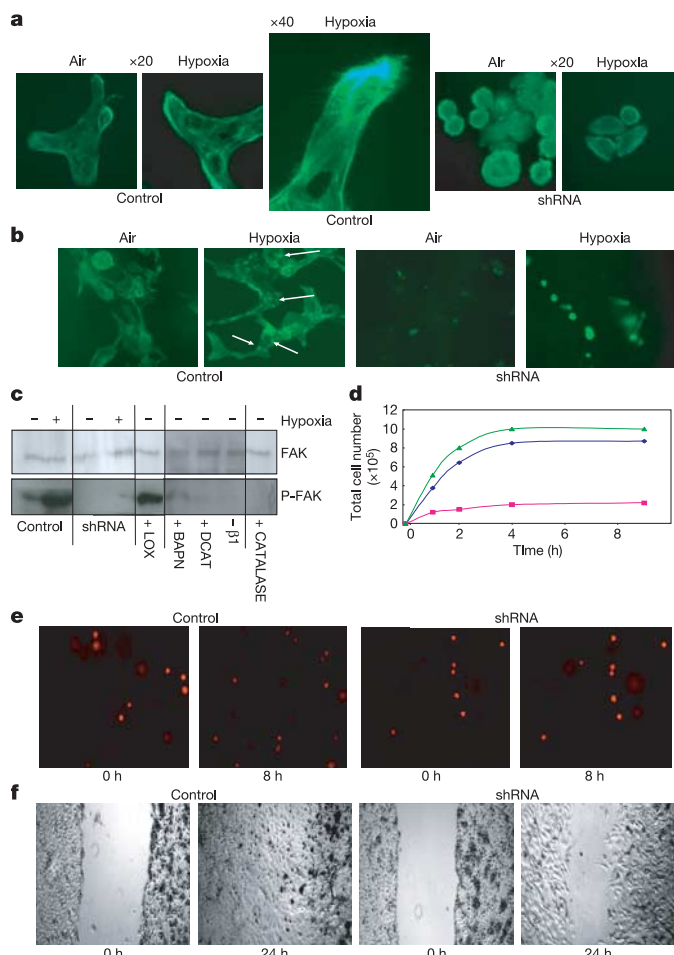


Figure 4 | LOX is required for adhesion interactions necessary for migration. **a**, LOX immunofluorescent staining of MDA231 cells. Note extracellular staining and intense staining at the leading edge of invasive pseudopod protrusions. Original magnifications are as shown. **b**, Phalloidin staining for F-actin in MDA231 cells. Arrows indicate examples of focal adhesions, confirmed by co-staining for FAK (data not shown). Exposure times (from left to right) were 1 min, 10 s, 1 min and 1 min. **c**, Immunoblot for the expression of total FAK and phospho-FAK in MDA231 cells (see Methods for details). **d**, Adhesion of MDA231 cells to matrix assessed over time. Blue, wild type; red, shRNA; green, shRNA+LOX. **e**, Time-lapse photography of control and LOX shRNA SiHa cells stably expressing DsRed in a collagen matrix. **f**, Time-lapse photography of a scratch assay performed with control and LOX shRNA MDA231 cells.

Elimination of the early invasive steps of metastasis through tail-vein studies revealed a role for LOX in the later stages of metastasis: mice injected with LOX shRNA cells had fewer lung foci (Supplementary Fig. 7a). Although the cloning efficiency of the cells was similar to that of control cells, their metastatic growth *in vitro* was severely impaired (Supplementary Fig. 7b, c). This seemed to be due to defective cell–matrix or extracellular matrix (ECM)–protein interactions required to permit and support abundant growth. Consistent with this is our observation that the metastatic lesions formed in our orthotopic studies were much larger in mice implanted with control cells than with shRNA cells, and were completely absent in BAPN-treated mice (Supplementary Fig. 7d). Close examination of these metastatic lesions revealed they were mostly composed of inflammatory cells. A recent report²² describes how factors secreted from tumour cells result in the recruitment of bone-marrow-derived cells (BMDCs), cluster formation and increased FN expression, providing a permissive niche for incoming tumour cells. Our data so far indicate that LOX might be a tumour-secreted factor required to create a permissive niche to support metastatic tumour cell growth. The role of LOX in BMDC recruitment is currently under investigation.

Several factors might explain why LOX is required for metastatic growth. First, LOX is required for FAK activity, which is known to mediate cell proliferation and survival. Second, LOX might regulate FN activity through FAK activation, providing a permissive niche to support metastatic tumour cell growth²². Last, LOX activity is essential for the formation of a mature ECM, which is undoubtedly required for survival signalling and cellular growth. It is noteworthy that we did not observe significant effects of LOX inhibition on primary tumour growth, whereas we found marked effects on metastatic growth in the lungs and liver. In particular, shRNA-expressing cells orthotopically implanted grew as primary tumours with the same kinetics as wild-type cells. These data indicate that the effects of LOX on cell adhesion, migration, invasion and three-dimensional growth are less crucial for primary tumour growth than for metastatic growth.

The data presented in this paper provide strong evidence that LOX is a good therapeutic target for the prevention of metastasis in breast cancer, and that targeting secreted LOX presents a mechanism for preventing early and late stages of metastasis. We have shown that hypoxia increases LOX mRNA, LOX protein, and secreted LOX activity, resulting in enhanced invasive migration required for metastatic spread. In addition, the remodelled matrix tracks resulting from increased LOX activity and cell migration could provide a highway along which other cells can travel more easily, thus increasing migration and invasion²³. Although LOX is known to be induced and/or activated by growth factors such as transforming growth factor- β ⁴, hypoxia might be more clinically relevant with regard to tumour progression. Many studies have shown that hypoxia promotes the aggressiveness of cancer cells². Whereas inhibition of LOX under aerobic conditions affected tumour cell invasion only modestly, inhibition of the much higher LOX expression levels observed under hypoxia consistently produced more marked effects. Furthermore, the aerobic conditions (21% O₂) typically employed for many *in vitro* assays⁹ are actually hyperoxic relative to oxygenation levels experienced in the body, particularly within solid tumours. Indeed, LOX expression was undetectable in some tumour cell lines without oxygen deprivation (Supplementary Fig. 1), again highlighting the relevance and clinical implications of hypoxia-induced levels of LOX.

We propose that hypoxia-induced LOX has a key function in tumour metastasis. Our data provide mechanistic evidence for hypoxia-driven metastasis and the influence of the ECM on metastatic spread, and support the therapeutic targeting of LOX to prevent and treat metastatic disease.

METHODS

Cell culture and oxygen deprivation. Cell lines obtained from ATCC were routinely cultured and oxygen deprived for 18 h as described²⁴. DsRed SiHa cells

were a gift from N. Dornhöfer. Actinomycin D (5 μ g ml⁻¹), desferoxamine/CoCl₂/bathocuproinedisulphonic acid/catalase (100 μ M) and BAPN (200 μ M) were used at given concentrations (Sigma).

Retrospective clinical study. The average hypoxia score in each breast cancer sample was calculated by averaging the expression value of all 122 unique unigene clusters comprising the hypoxia gene signature without LOX, as described previously¹². The correlation between the averaged hypoxia score and LOX expression in each breast cancer sample was then calculated. For Kaplan–Meyer plots, LOX expression levels were determined as described²⁵, plotting top and bottom tiers.

Immunology studies. Methods were as described previously²⁴. Primary antibodies used were as follows: HIF-1 α (BD Biosciences), α -tubulin (Research Diagnostics), FAK (Sigma) and phospho-FAK (Santa Cruz). LOX polyclonal antibody was raised against a synthetic peptide of human LOX (EDTSCDYGYYHRRFA; Open Biosystems). Hypoxic regions were identified by staining for Pimonidazole²⁴. For LOX immunoblotting, proteins in CM were concentrated with Microcon filters (Millipore), and 20 μ l portions were loaded. Quantification was performed with ImageQuant software.

Semiquantitative and fully quantitative reverse transcriptase polymerase chain reaction analysis. Methods were as described²⁴. The Taqman polymerase chain reaction primer sequences for LOX were 5'-ATGAGTTTAGCCACTTG TACCTGCTT-3' and 5'-AACTTGCTTGTGGCCTTCA-3'.

Hypoxia responsiveness of the LOX promoter. The luciferase assay was conducted as described²⁴. LOX promoter sequence was cloned into pGL3-Basic (Promega) and mutated by site-directed mutagenesis (Stratagene), in accordance with the manufacturer's instructions. A positive control containing five hypoxia-responsive elements (HREs) was used²⁶.

Transient and retroviral transfections. Treatment of cells with sense or antisense phosphorothioate-modified LOX-specific sense or antisense oligonucleotides (Integrated DNA Technologies) was as described^{8,24}. HIF-1 oxygen-dependent degradation (ODD) domain and proline double mutant constructs were a gift from D. Chan. Mature LOX and the LOX delta cat mutant were gifts from A. Di Donato. HIF-1 α shRNA was as described²⁴. All transient transfections were performed with Lipofectamine 2000 (Invitrogen) in accordance with the manufacturer's instructions. Transfection efficiencies were about 90% for SiHa cells, and about 50% for MDA231 and 435 cells.

MDA-MB 231 and SiHa human cancer cells were retrovirally transfected with pBabe vector²⁷ containing a LOX-specific targeting sequence (5'-GTTCCTGC TCTCAGTAACC-3'). This sequence was checked against the database to confirm specificity. A scrambled sequence was used as a control (5'-CACATGTTCCG ATCTCGGC-3'). Infected cells were selected in puromycin (Sigma); polyclonal cell populations were then tested for decreased LOX expression levels.

***In vitro* invasion analysis.** The method used was as described²⁸. In brief, cells were serum deprived for 24 h, then 2.5×10^4 cells were seeded in triplicate on both Matrigel-coated and uncoated inserts, moved to chambers containing 750 μ l of 10% FBS as a chemo-attractant and incubated under normoxic or oxygen-deprived conditions for 24 h. Treatments with BAPN/BCS were performed 24 h before serum deprivation, and continued throughout the experiment.

Assessment of growth of cells in culture. For monolayer growth curves, cells were counted each day with a haemocytometer. Three-dimensional growth in Matrigel and type I collagen (both from BD Biosciences) was performed as described¹⁶. In brief, 2×10^4 cells were seeded and grown for three days before a four-day incubation (collagen), or simply for ten days (Matrigel) under normoxia or hypoxia.

***In vivo* tail-vein metastasis assay.** Control and LOX shRNA MDA231 cells were injected intravenously with 5×10^5 cells in 0.1 ml of DMEM medium into the tail vein. Four weeks after injection, mice were killed. Microscopic quantification of lung foci was performed on representative cross-sections of formalin-fixed, paraffin-embedded lungs stained with haematoxylin and eosin. The correct identification of lung foci (minimum of four human cells with large nuclei) was kindly confirmed by a board-certified veterinary pathologist.

Growth of MDA-MB 231 cells as orthotopic tumours. MDA-MB 231 cells were grown as subcutaneous orthotopic tumours in six-week-old female Nude (*nu/nu*) mice after intradermal injection of 10^7 cells in 0.1 ml of PBS into the mammary fat pad. Mice were killed and tumours were excised six weeks after inoculation. Some mice were treated twice weekly with up to 20 mg kg⁻¹ purified LOX antibody, two weeks after inoculation, or daily with 100 mg kg⁻¹ BAPN (intraperitoneally, as described²⁹) for the last four, three or two weeks of tumour growth (4, 3 and 2 wk BAPN, respectively; these doses and durations have no deleterious side effects²⁹). Lung and liver metastases were defined as gross lesions of at least 25 cells. The presence of human tumour cells was verified by cytokeratin staining (not shown).

LOX activity assay. The original method is described in ref. 19. For the *in vivo* data, terminal bleeds were taken at the end of the experiment described above from untreated (control) mice, 2 and 3 wk BAPN mice, and antibody-treated mice. Plasma (10 μ l) was tested and fluorescence (a measure of LOX activity) was plotted, where 0 = sample + 500 μ M BAPN (complete LOX inhibition). For the *in vitro* data, 50 μ l of conditioned phenol-red-free medium was taken from cells incubated for 24 h under conditions of hypoxia. Samples were incubated overnight at 37 °C with different concentrations of BAPN or with purified LOX antibody at a concentration equivalent to 20 mg kg⁻¹ dose. Again, 0 = fluorescent reading for 500 μ M BAPN. No activity could be detected in CM from aerobic and LOX shRNA-expressing cells, or in blood from mice that did not bear tumours (data not shown).

Adhesion assay. For MDA231 cells, 2.5×10^5 cells were plated in serum-free medium on collagen-coated 96-well plates. Adherent cells were trypsinized and counted over an 8-h time course.

For DSRED SiHa cells, 2.5×10^5 cells were plated in serum-free medium and left to adhere. Wells were washed and medium was replaced with PBS at specific times over an 8-h time course, after which fluorescence was measured and the number of adherent cells was determined from a standard curve (generated with 2.5×10^3 to 5×10^5 cells).

Received 13 January; accepted 2 March 2006.

- Cairns, R. A., Khokha, R. & Hill, R. P. Molecular mechanisms of tumor invasion and metastasis: An integrated view. *Curr. Mol. Med.* **3**, 659–671 (2003).
- Hockel, M. & Vaupel, P. Tumor hypoxia: Definitions and current clinical, biologic, and molecular aspects. *J. Natl Cancer Inst.* **93**, 266–276 (2001).
- Denko, N. C. *et al.* Investigating hypoxic tumor physiology through gene expression patterns. *Oncogene* **22**, 5907–5914 (2003).
- Kagan, H. M. & Li, W. Lysyl oxidase: Properties, specificity, and biological roles inside and outside of the cell. *J. Cell. Biochem.* **88**, 660–672 (2003).
- Kaneda, A. *et al.* Lysyl oxidase is a tumor suppressor gene inactivated by methylation and loss of heterozygosity in human gastric cancers. *Cancer Res.* **64**, 6410–6415 (2004).
- Csiszar, K. *et al.* Somatic mutations of the lysyl oxidase gene on chromosome 5q23.1 in colorectal tumors. *Int. J. Cancer* **97**, 636–642 (2002).
- Palamakumbura, A. H. *et al.* The propeptide domain of lysyl oxidase induces phenotypic reversion of ras-transformed cells. *J. Biol. Chem.* **279**, 40593–40600 (2004).
- Kirschmann, D. A. *et al.* A molecular role for lysyl oxidase in breast cancer invasion. *Cancer Res.* **62**, 4478–4483 (2002).
- Payne, S. L. *et al.* Lysyl oxidase regulates breast cancer cell migration and adhesion through a hydrogen peroxide-mediated mechanism. *Cancer Res.* **65**, 11429–11436 (2005).
- Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl Acad. Sci. USA* **102**, 3738–3743 (2005).
- Chen, Y. *et al.* Identification of hypoxia-regulated proteins in head and neck cancer by proteomic and tissue array profiling. *Cancer Res.* **64**, 7302–7310 (2004).
- Chi, J. T. *et al.* Gene expression programs in response to hypoxia: Cell type specificity and prognostic significance in human cancers. *PLoS Med.* **3**, e47 (2006).
- Palmieri, D. *et al.* Medroxyprogesterone acetate elevation of Nm23-H1 metastasis suppressor expression in hormone receptor-negative breast cancer. *J. Natl Cancer Inst.* **97**, 632–642 (2005).
- Csiszar, K. Lysyl oxidases: a novel multifunctional amine oxidase family. *Prog. Nucleic Acid Res. Mol. Biol.* **70**, 1–32 (2001).
- Cen, D., Brayton, D., Shahandeh, B., Meyskens, F. L. Jr & Farmer, P. J. Disulfiram facilitates intracellular Cu uptake and induces apoptosis in human melanoma cells. *J. Med. Chem.* **47**, 6914–6920 (2004).
- Janda, E. *et al.* Ras and TGF β cooperatively regulate epithelial cell plasticity and metastasis: dissection of Ras signaling pathways. *J. Cell Biol.* **156**, 299–313 (2002).
- Friedl, P. & Wolf, K. Tumour-cell invasion and migration: diversity and escape mechanisms. *Nature Rev. Cancer* **3**, 362–374 (2003).
- Krishnamachary, B. *et al.* Regulation of colon carcinoma cell invasion by hypoxia-inducible factor 1. *Cancer Res.* **63**, 1138–1143 (2003).
- Fogelgren, B. *et al.* Cellular fibronectin binds to lysyl oxidase with high affinity and is critical for its proteolytic activation. *J. Biol. Chem.* **280**, 24690–24697 (2005).
- Yoon, S. O., Shin, S. & Mercurio, A. M. Hypoxia stimulates carcinoma invasion by stabilizing microtubules and promoting the Rab11 trafficking of the α 6 β 4 integrin. *Cancer Res.* **65**, 2761–2769 (2005).
- Leufgen, H. *et al.* Collagenase expression and activity is modulated by the interaction of collagen types, hypoxia, and nutrition in human lung cells. *J. Cell. Physiol.* **204**, 146–154 (2005).
- Kaplan, R. N. *et al.* VEGFR1-positive haematopoietic bone marrow progenitors initiate the pre-metastatic niche. *Nature* **438**, 820–827 (2005).
- Wang, W. *et al.* Gene expression analysis on small numbers of invasive cells collected by chemotaxis from primary mammary tumors of the mouse. *BMC Biotechnol.* **3**, 13 (2003).
- Erler, J. T. *et al.* Hypoxia-mediated down-regulation of Bid and Bax in tumors occurs via hypoxia-inducible factor 1-dependent and -independent mechanisms and contributes to drug resistance. *Mol. Cell. Biol.* **24**, 2875–2889 (2004).
- Moody, S. E. *et al.* The transcriptional repressor Snail promotes mammary tumor recurrence. *Cancer Cell* **8**, 197–209 (2005).
- Tazuke, S. I. *et al.* Hypoxia stimulates insulin-like growth factor binding protein 1 (IGFBP-1) gene expression in HepG2 cells: a possible model for IGFBP-1 expression in fetal hypoxia. *Proc. Natl Acad. Sci. USA* **95**, 10188–10193 (1998).
- Spicher, A. *et al.* Highly conserved RNA sequences that are sensors of environmental stress. *Mol. Cell. Biol.* **18**, 7371–7382 (1998).
- Bedogni, B. *et al.* Topical treatment with inhibitors of the phosphatidylinositol 3'-kinase/Akt and Raf/mitogen-activated protein kinase/extracellular signal-regulated kinase pathways reduces melanoma development in severe combined immunodeficient mice. *Cancer Res.* **64**, 2552–2560 (2004).
- Gilad, G. M. & Gilad, V. H. Beta-aminopropionitrile treatment can accelerate recovery of mice after spinal cord injury. *Eur. J. Pharmacol.* **430**, 69–72 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Chu (immunohistochemistry), D. Menke (tail-vein injections), R. Nacamuli (quantitative polymerase chain reaction), R. R. Balise (statistical programming) and C. Davis (board-certified veterinary pathologist); N. Quach (FAK, phospho-FAK and F-actin antibodies), A. Cress (α_6 antibody), P. Marinkovich (β_1 integrin antibody) and D. Mosher (FN-null cells) for supplying reagents; and P. Friedl, Z. Werb and P. Steeg for discussions. All animal work was performed in accordance with the Stanford University Administrative Panel for Laboratory Animal Care. This research was supported by funds from the NIH (J.T.E. and A.J.G.), the Canadian Institutes of Health Research (K.L.B.), the Else Kröner-Fresenius-Foundation (N.D.) and the California Breast Cancer Research Program (S.S.J.).

Author Contributions J.T.E., A.J.G. and S.S.J. conceived and designed the experiments. J.T.E. performed the experiments with the assistance of K.L.B. and N.D. for *in vivo* work. J.T.E., A.J.G., M.N., J.T.C. and S.S.J. analyzed the data. J.T.E. wrote the paper with the assistance of K.L.B., S.S.J. and A.J.G. The head and neck cancer study was performed by Q.T.L. and C.K.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.J.G. (giaccia@stanford.edu).

Detecting transient intermediates in macromolecular binding by paramagnetic NMR

Junji Iwahara¹ & G. Marius Clore¹

Macromolecular complex formation is governed by two opposing constraints of specificity and speed^{1,2}. Kinetic^{3–6} and theoretical considerations suggest that significant rate enhancement can be achieved either by reducing the dimensionality of the search process^{1,7} or by the creation of a short-range attractive potential around the target site². This implies the existence of transient intermediates involving non-specific binding modes. Here we show that intermolecular paramagnetic relaxation enhancement (PRE) provides a means of directly detecting the presence of, and investigating the nature of, low population transient intermediates under equilibrium conditions. Applying this approach, we characterize the search process whereby a sequence-specific transcription factor (the homeodomain of HOXD9) binds to non-cognate DNA sites as a means of enhancing the rate of specific association. The PRE data in the fast exchange regime reveal the presence of transient intermediates formed in a stochastic manner at non-cognate sites whose structure is similar to that of the specific complex. Two distinct search processes involving intra- as well as intermolecular translocations can be delineated. The intermolecular PRE method is general and can be readily applied to investigations of transient intermediates in many other macromolecular binding processes.

The PRE is caused by magnetic dipolar interactions between a nucleus and the unpaired electrons of a paramagnetic centre, and results in an increase in the relaxation rate of the nuclear magnetization^{8,9}. The magnitude of the PRE is proportional to $\langle r^{-6} \rangle$ (where r is the distance between the nucleus of interest and the paramagnetic centre) and, owing to the large magnetic moment of an unpaired electron, the effect is detectable for sizeable separations (up to ~ 34 Å for Mn^{2+}). The impact on the transverse PRE, Γ_2 , observed on the resonance of a major species A (99% occupancy) in a two-site exchange system comprising a minor species B (1% occupancy) with corresponding paramagnetic centre–proton distances of 30 Å and 8 Å, respectively, is shown in Fig. 1. (Γ_2 is defined as the difference in the transverse relaxation rates of the paramagnetic and diamagnetic states.) For a ~ 30 -kilodalton (30-kDa) complex, the ^1H - Γ_2 arising from Mn^{2+} is $\sim 2\text{ s}^{-1}$ ($\Gamma_{2,A}$) for species A but $\sim 5.6 \times 10^3\text{ s}^{-1}$ ($\Gamma_{2,B}$) for species B¹⁰. The apparent value of Γ_2 (Γ_2^{app}) is highly dependent on the exchange rate, k_{ex} , between the two species. If k_{ex} is slow ($< 50\text{ s}^{-1}$), the presence of the minor species B has no effect and the value of Γ_2^{app} is the same as that expected for the major species A. For larger k_{ex} , however, Γ_2^{app} is highly influenced by the minor species B. When $k_{\text{ex}} \gg (\Gamma_{2,B} - \Gamma_{2,A})$, Γ_2^{app} is the weighted population average of the Γ_2 rates in the two species¹¹. Under these conditions Γ_2^{app} is ~ 30 -fold larger than $\Gamma_{2,A}$, thereby permitting one to both infer the presence of, and obtain some structural information on, a minor species.

We studied the interaction between the homeodomain of human HOXD9 and a 24-base-pair (24-bp) DNA duplex containing a single specific binding site (Fig. 2a). Conjugated deoxy (d)T-EDTA- Mn^{2+}

(ref. 12; EDTA, ethylene diamine tetra-acetic acid) was placed at four distinct sites (one per duplex), two (sites 1 and 4) close to the ends of the DNA, and two (sites 2 and 3) adjacent to the 5' and 3' ends of the specific target site (Fig. 2a). Homeodomains are found in many eukaryotic transcription factors and possess well-characterized sequence-specific DNA-binding activity¹³. Because of the high degree of sequence identity to the *Drosophila* Antp homeodomain/DNA complex studied previously by both NMR¹⁴ and crystallography¹⁵, as well as the excellent agreement between measured dipolar couplings and those calculated from the related crystal structure¹⁵ (dipolar coupling R -factor¹⁶ of 14–15%, see Supplementary Information), the structure of the HOXD9 homeodomain/DNA complex can be readily

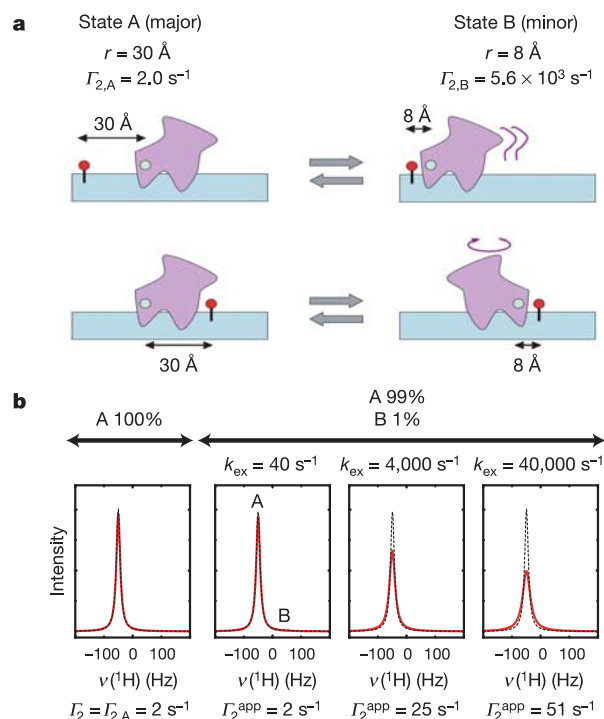


Figure 1 | Intermolecular PRE in an exchanging system. **a**, Exchange between two states A (major, 99%) and B (minor, 1%) in which the distance, r , from the monitored proton (green circle) on the protein (lilac) to the paramagnetic centre (red ball) on the DNA (cyan) is 30 Å and 8 Å, respectively. The transition between the two states may involve either translation (top) or rotation (bottom) of the protein relative to the DNA. **b**, Simulations based on the McConnell equation²⁷. The solid red lines and dotted black lines represent NMR line-shapes with and without PRE, respectively (for details, see Supplementary Information). The definitions of Γ_2 , Γ_2^{app} , $\Gamma_{2,A}$, $\Gamma_{2,B}$ and k_{ex} are given in the main text.

¹Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA.

modelled. From two-dimensional ^{15}N -exchange spectroscopy, the apparent overall exchange rate for intermolecular translocation of the HOXD9 homeodomain between DNA specific sites (located on two DNA duplexes differing by only a single base pair just outside the specific binding site) is slow ($k_{\text{ex}} = 7 \text{ s}^{-1}$) on the NMR chemical shift timescale at 20 mM NaCl but fast ($k_{\text{ex}} \approx 600 \text{ s}^{-1}$) at 160 mM NaCl, under conditions where the concentration of total free DNA is $\sim 200 \mu\text{M}$ (see Supplementary Information). The equilibrium dissociation constant (K_{diss}) measured by fluorescence anisotropy is 1.5 nM at 100 mM NaCl and 10 nM at 160 mM NaCl (Supplementary Information). $^1\text{H}_\text{N}$ - Γ_2 PREs were measured for the backbone amide groups ($^1\text{H}_\text{N}$) of complexed $^2\text{H}/^{15}\text{N}$ -labelled HOXD9, and were found to be completely different in the slow- (Fig. 2b, 20 mM NaCl) and fast- (Fig. 2c, 100 and 160 mM NaCl) exchange regimes for all four dT-EDTA- Mn^{2+} sites (see Supplementary Information).

Under the experimental conditions employed, involving relatively high (submillimolar) concentrations of free DNA, the exchange process observed in the ^{15}N -exchange experiments does not involve spontaneous dissociation of the protein followed by reassociation of free protein and DNA, but occurs via direct transfer following collision of free DNA with DNA-bound protein to form a transient ternary encounter complex without ever going through the intermediary of free protein¹⁷. This mechanism, which is akin to inter-segment transfer³, dramatically accelerates the rate of target recognition in protein–DNA interactions, resulting in translocation rates that are over 3 orders of magnitude faster than the dissociation

rate constant determined by gel shift assays at very low (nM) concentrations of free DNA ($k_{\text{diss}} \ll 0.01 \text{ s}^{-1}$ for the HOXD9 homeodomain)^{18,19}. This reconciles the highly dynamic behaviour of protein–DNA complexes observed *in vivo* using microscopy combined with photobleaching techniques²⁰ with the long half-lives of specific protein–DNA complexes measured by traditional biochemical analysis *in vitro*¹⁷.

At 20 mM NaCl, the $^1\text{H}_\text{N}$ - Γ_2 data arising from the four dT-EDTA- Mn^{2+} groups are fully consistent with the structure of the complex bound to the specific site (Figs 2b and 3c; see also Supplementary Information). Large magnitude $^1\text{H}_\text{N}$ - Γ_2 PREs ($>10 \text{ s}^{-1}$) are only observed for those regions in relatively close proximity to the dT-EDTA- Mn^{2+} groups, and the observed $^1\text{H}_\text{N}$ - Γ_2 values are in excellent agreement with those predicted from the model of the specific complex with an overall PRE Q-factor¹⁰ of 0.26 (Fig. 2d). Thus, in the slow exchange regime, the presence of intermediate states, as expected from the lineshape simulations shown in Fig. 1, is not apparent from the PRE data.

At 100 and 160 mM NaCl, however, many residues exhibit $^1\text{H}_\text{N}$ - Γ_2 PREs that are completely inconsistent with the structure of the specific complex (Figs 2c, e and 3d, and Supplementary Fig. S3).

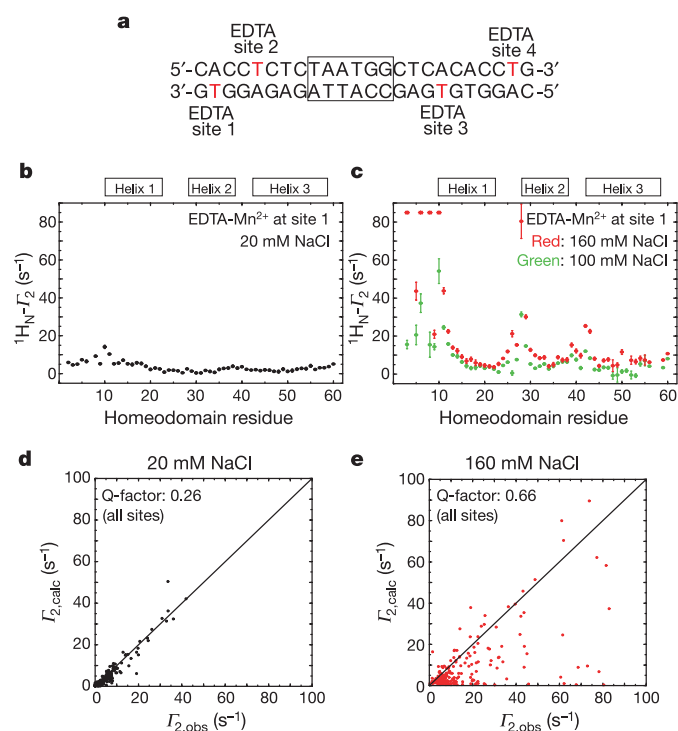


Figure 2 | Intermolecular PRE for the HOXD9 homeodomain/DNA complex in slow and fast exchange. **a**, 24-bp duplex DNA. The specific target site (TAATGG) is boxed. The four sites chosen to covalently attach EDTA to thymine (one site at a time) are shown in red. **b**, **c**, Intermolecular PRE $^1\text{H}_\text{N}$ - Γ_2 profiles obtained for EDTA- Mn^{2+} at site 1 in 20 mM (**b**), 100 mM (**c**, green) and 160 mM (**c**, red) NaCl. Asterisks indicate residues whose $^1\text{H}_\text{N}$ - ^{15}N cross-peaks are broadened beyond the limits of detection. Γ_2 profiles for sites 2–4 are shown in Supplementary Fig. S3. **d**, **e**, Correlation between observed and calculated PREs at all 4 sites in 20 mM (**d**) and 160 mM (**e**) NaCl (217 and 197 data points, respectively). The PRE Q-factor¹⁰ is a measure of agreement between observed and calculated values of Γ_2 and is given by $[\sum_i (\Gamma_2^{\text{obs}}(i) - \Gamma_2^{\text{calc}}(i))^2 / \sum_i \Gamma_2^{\text{obs}}(i)^2]^{1/2}$. Error bars, $\pm 1 \text{ s.d.}$

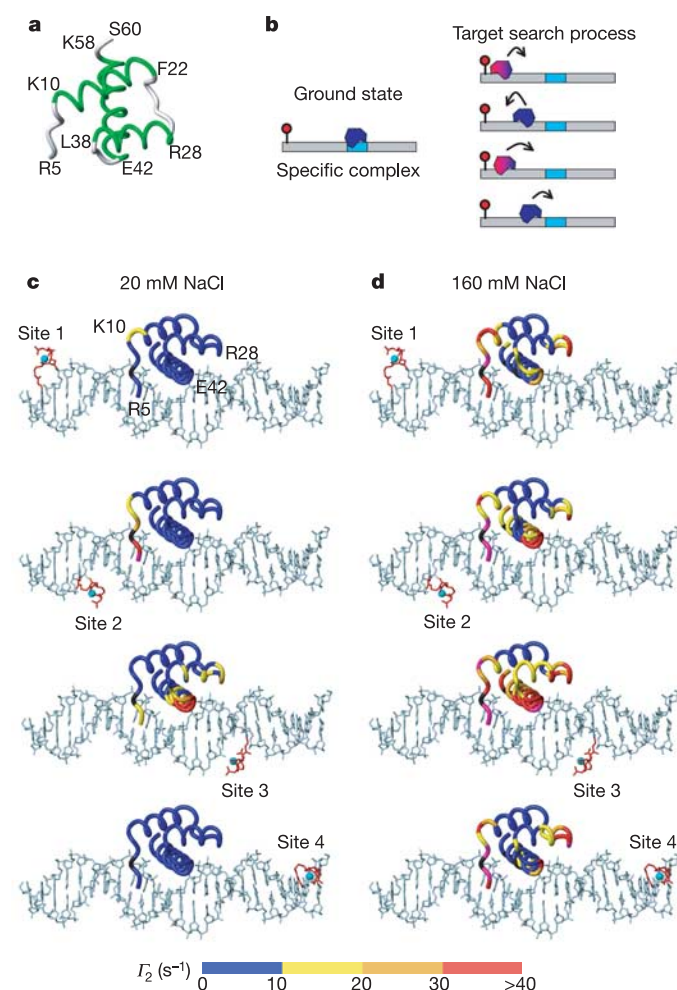


Figure 3 | Summary of the intermolecular PRE profiles arising from dT-EDTA- Mn^{2+} at sites 1 to 4 for the HOXD9 homeodomain/DNA complex in the slow and fast exchange regimes. **a**, Ribbon diagram of the HOXD9 homeodomain. **b**, Schematic representation of the specific target search process involving non-specific binding intermediates that experience strong intermolecular PRE (indicated by the colour gradient from red to blue, with red indicating strong PRE). **c**, **d**, Intermolecular PRE $^1\text{H}_\text{N}$ - Γ_2 data from sites 1 to 4 at 20 mM (**c**) and 160 mM (**d**) NaCl mapped on the structural model of the HOXD9/DNA complex: colour scale shows Γ_2 .

For example, Arg 28 and Asp 29 (at the amino terminus of helix 2) exhibit $^1\text{H}_\text{N}$ - Γ_2 values $>30\text{ s}^{-1}$ but in the specific complex are located on the opposite face of the protein with respect to dT-EDTA-Mn $^{2+}$ at sites 1 and 2 with corresponding Mn $^{2+}$ - $^1\text{H}_\text{N}$ distances $>40\text{ \AA}$. The carboxy-terminal half of helix 1 (Glu 15 to Asn 23) is closer to dT-EDTA-Mn $^{2+}$ at sites 1 and 2 than Arg 28 and Asp 29, yet displays $^1\text{H}_\text{N}$ - Γ_2 values $<10\text{ s}^{-1}$ that are minimally affected by salt concentration. The dramatic changes in $^1\text{H}_\text{N}$ - Γ_2 PRE profiles upon increasing the ionic strength from 20 to 160 mM NaCl are not due to any significant structural changes in the specific complex, because (1) the ^1H - ^{15}N correlation spectrum of the HOXD9 homeodomain bound to DNA is essentially identical at 20 and 160 mM NaCl, and (2) the residual dipolar couplings at the two salt conditions are highly correlated (correlation coefficient, 0.99; see Supplementary Fig. S2). We therefore conclude that the intermolecular PRE data at 100 and 160 mM NaCl represent the footprint of minor species that exchange rapidly with the specific complex. The HOXD9 homeodomain in these minor states is bound in a stochastic manner to various sites on the DNA, and can be located in close spatial proximity to the dT-EDTA-Mn $^{2+}$ sites. The population of the minor species is less than 1%, judging from the values of the specific and non-specific equilibrium dissociation constants (1.5 versus 270 nM at 100 mM NaCl).

The overall PRE profiles at 160 mM NaCl also provide structural information relating to non-specific DNA binding of the HOXD9 homeodomain. Patches on the protein with large PREs (Fig. 3d) indicate that these regions can come into close proximity to conjugated dT-EDTA-Mn $^{2+}$ located in the DNA major groove, while regions with small PREs, such as the C-terminal half of helix 1 (Glu 15–Asn 23), must be distant from the DNA interface even in the non-specific complexes. The PRE map therefore suggests that the DNA binding mode adopted during the target search process is similar to that in the specific complex, and that the populations of any potential species involving alternative protein interaction surfaces, should these exist, are below the limits of detection.

The target search process can involve either intramolecular translocation of the protein (for example, sliding or one-dimensional diffusion along the DNA) or intermolecular translocation in which the protein is directly transferred from one DNA molecule to another (Fig. 4a). To evaluate their relative contributions, we carried out additional PRE experiments on two samples comprising an equal mixture of two DNA duplexes, one with and the other without the specific target site. EDTA-Mn $^{2+}$ was conjugated to the non-specific DNA duplex in sample 1 and to the specific DNA duplex in sample 2 (Fig. 4b). Thus, for sample 1 the intermolecular PRE can only arise through intermolecular translocation, while for sample 2 both intra- and intermolecular processes can contribute. The ^1H - ^{15}N correlation spectrum of HOXD9 in both samples was basically the same as that of the specific complex, indicating that essentially all protein is bound to the specific target site (see Supplementary Fig. S4). The PRE profiles observed for the two samples at 160 mM NaCl (Fig. 4c) are similar, but the magnitude of the PREs for sample 2 are generally larger than those for sample 1, indicating that both intra- and intermolecular translocations are involved in the target search. From the ratio of the Γ_2 values for samples 1 and 2 (Fig. 4c), the contribution of intermolecular translocation is actually larger than that of intramolecular translocation. This is due to the high total concentration of free DNA (0.8 mM) in the NMR samples. As the concentration of DNA in the nucleus is as high as $\sim 100\text{ mg ml}^{-1}$ (equivalent to 150 mM on a per base pair basis)²¹, it seems quite feasible that intermolecular translocation (or intersegment transfer) could contribute significantly to the speed of the search process *in vivo*¹⁷. Quantitative analysis of the PRE data for samples 1 and 2 also reveals the existence of one-dimensional sliding along the DNA. For Met 24–Glu 33 and Thr 41–Glu 42, the PREs measured for sample 2 are systematically larger by ~ 30 –100% than those for sample 1, whereas those for the N-terminal arm region are almost

the same for samples 1 and 2 (Fig. 4c). These observations can be attributed to bias in which the orientation of the protein bound to the specific site is favoured as the protein slides (intra-molecular translocation) along DNA. Thus, in the sliding process the segments Met 24–Glu 33 and Thr 41–Glu 42 can come very close to the EDTA-Mn $^{2+}$ on the oligonucleotide containing the specific target site in sample 2 (DNA 2A in Fig. 4b) whereas the N-terminal arm cannot be so close unless the protein takes up the opposite orientation on the DNA, which is allowed by intermolecular translocation (see Fig. 4d).

In conclusion, intermolecular PRE data for complexes in the fast exchange regime can provide valuable information on intermediates

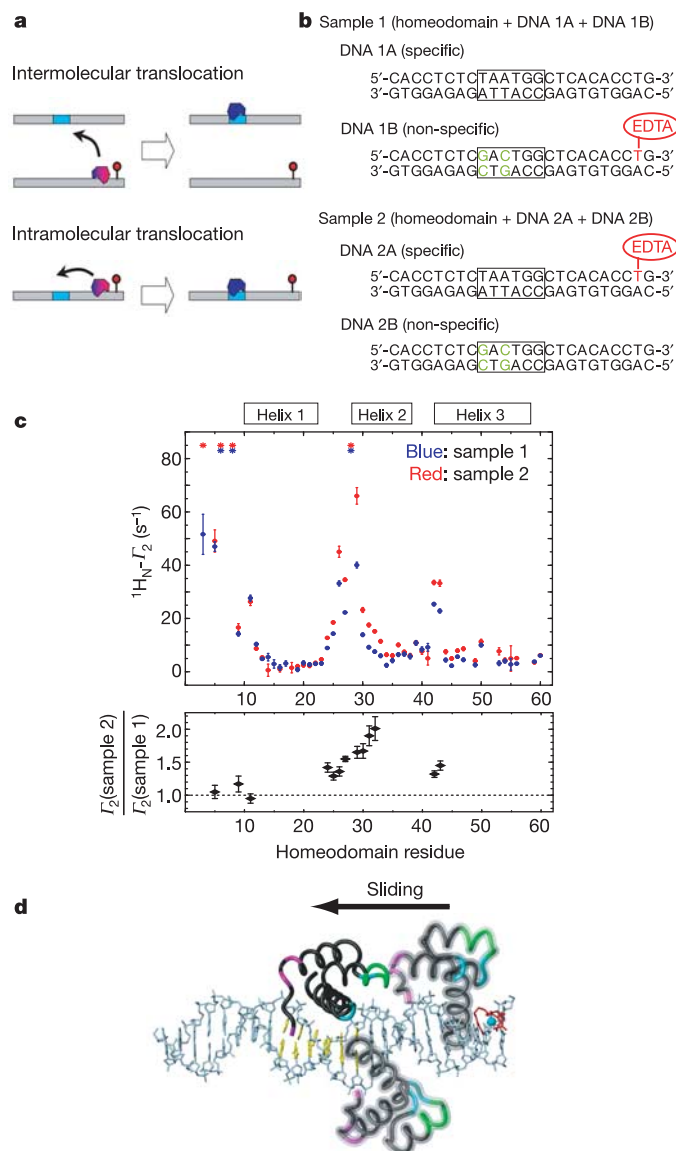


Figure 4 | Intermolecular PRE arises from inter- and intramolecular translocation processes. **a**, Diagrammatic representation of inter- and intramolecular translocation giving rise to observable PREs. **b**, DNA samples. **c**, PRE data for samples 1 (blue) and 2 (red) are displayed on the top panel (asterisks represent cross-peaks broadened beyond the limits of detection); the bottom panel shows the $\Gamma_2(\text{sample 2})/\Gamma_2(\text{sample 1})$ ratios for residues with $\Gamma_2 > 12\text{ s}^{-1}$ in sample 2. **d**, Schematic representation of sliding along the DNA: homeodomain residues with $\Gamma_2(\text{sample 2})/\Gamma_2(\text{sample 1})$ ratio <1.25 , magenta; 1.25 to 1.5, cyan; and >1.5 , green. The specific DNA target site is coloured in yellow. Specifically and non-specifically (two) bound homeodomains are depicted as opaque and transparent tubes, respectively. Error bars, $\pm 1\text{ s.d.}$

whose populations at equilibrium are very low. These measurements not only reveal the presence of intermediate states but also shed light on their structural features. Although the methodology was applied to a protein–DNA complex, a similar approach can be equally well applied to protein–protein complexes (unpublished data), as extrinsic paramagnetic groups can be readily linked covalently to proteins via surface cysteine residues (selectively introduced by mutagenesis)^{22,23}. Experimental data on the detection of low populations of intermediates at equilibrium, together with appropriate computational studies, such as brownian dynamics simulations²⁴, should yield a deeper insight into macromolecular recognition processes.

METHODS

Sample preparation. The human HOXD9 homeodomain was prepared as described¹⁷. 24-bp DNA fragments with and without dT-EDTA were prepared as described^{10–12}. To make metal-free samples of DNA, free EDTA (50 mM) was initially added and then removed by exchanging the buffer to 20 mM Tris HCl (pH 6.8), 500 mM NaCl, using an Amicon Ultra spin concentrator. The EDTA-conjugated DNA in the apo-state was mixed with metal ions (Mn^{2+} or Ca^{2+}) in ~50% excess relative to DNA. The HOXD9 homeodomain was added to the mixture at a protein-to-DNA ratio of 1:1.5 in the presence of 500 mM NaCl. The excess metal ions located at sites other than the conjugated EDTA were removed by extensive washing with 20 mM Tris HCl (pH 6.8) and 500 mM NaCl using a spin concentrator. After washing, the buffer for the HOXD9/DNA-EDTA-metal complexes was exchanged to 10 mM Tris HCl (pH 6.8), 20 mM NaCl and 7% D_2O for NMR measurements. All buffers were treated with chelex-20 (Sigma) to avoid contamination of metal ions.

NMR spectroscopy. 1H -, ^{13}C - and ^{15}N -resonances were assigned as described¹⁷. PRE 1H_N - T_2 data at a 1H frequency of 600 MHz were acquired at 35 °C using Bruker DRX-600 spectrometers equipped with cryogenic triple resonance probes. For individual PRE experiments, data on two samples comprising 0.4 mM $^2H/^{15}N$ -labelled HOXD9 homeodomain and 0.6 mM DNA-EDTA chelating either Ca^{2+} or Mn^{2+} were recorded, using two-dimensional 1H - ^{15}N correlation spectra described previously^{12,25}. The protein NMR chemical shifts for the Mn^{2+} and Ca^{2+} chelated states are identical as the g-tensor for Mn^{2+} is isotropic. Two time points with a difference of 14 ms were used for PRE 1H_N - T_2 measurements. Values of 1H_N - T_2 were calculated as described¹⁰. For the complex comprising dT-EDTA at site 4, the 1H -relaxation rates were also measured with eight 1H relaxation time points and the values of 1H_N - T_2 obtained were identical to those from the data recorded at two time points within experimental errors. The maximum value of T_2 that can be determined accurately is ~80–90 s⁻¹; beyond that, the 1H_N - ^{15}N cross-peaks are too broad to permit quantitative determination of the PRE.

Two samples were employed to analyse the contribution of intra- and intermolecular translocation processes to the intermolecular PRE (Fig. 4b). Sample 1 comprised $^2H/^{15}N$ -labelled HOXD9 homeodomain, DNA 1A (containing the specific target site but no EDTA-conjugation), and DNA 1B (with EDTA-conjugation at a position corresponding to site 4 and the specific target site removed by two base pair mutations). Sample 2 contained $^2H/^{15}N$ -labelled HOXD9 homeodomain, DNA 2A (containing the specific DNA target site and EDTA- Mn^{2+} conjugation at site 4) and DNA 2B (without specific target site and no EDTA- Mn^{2+} conjugation). The only difference between samples 1 and 2 is the location of the conjugated EDTA- Mn^{2+} (that is, on the DNA containing the specific target site in sample 2 and on the DNA with the specific target site removed in sample 1). The ratio of protein:DNA_{specific}:DNA_{non-specific} was 1:1.5:1.5 with a protein concentration of 0.4 mM.

Back-calculation of PREs. PREs were back-calculated from the structural model of the HOXD9/DNA complex using a three-conformer ensemble representation for the EDTA- Mn^{2+} groups to account for their flexibility¹⁰. The coordinates of the EDTA- Mn^{2+} moieties were optimized by simulated annealing using Xplor-NIH²⁶ as described previously¹⁰.

Received 28 November 2005; accepted 22 February 2006.

1. Berg, O. G. & von Hippel, P. H. Diffusion-controlled macromolecular interactions. *Annu. Rev. Biophys. Chem.* **14**, 131–160 (1985).
2. Zhou, H.-X. & Szabo, A. Enhancement of association rates by nonspecific binding to DNA and cell membranes. *Phys. Rev. Lett.* **93**, 178101 (2004).

3. von Hippel, P. H. & Berg, O. G. Facilitated target location in biological systems. *J. Biol. Chem.* **264**, 675–678 (1989).
4. Halford, S. E. & Marko, J. F. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.* **32**, 3040–3052 (2004).
5. Schreiber, G. & Fersht, A. R. Rapid, electrostatically assisted association of proteins. *Nature Struct. Biol.* **3**, 427–431 (1996).
6. Vijayakumar, M. et al. Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on barnase and barstar. *J. Mol. Biol.* **278**, 1015–1024 (1998).
7. Adam, G. & Delbruck, M. in *Structural Chemistry and Molecular Biology* (eds Rich, A. & Davidson, N.) 198–215 (Freeman & Co., San Francisco, 1968).
8. Solomon, I. Relaxation processes in a system of two spins. *Phys. Rev.* **99**, 559–565 (1955).
9. Bloembergen, N. & Morgan, L. O. Proton relaxation times in paramagnetic solutions. Effects of electron spin relaxation. *J. Chem. Phys.* **34**, 842–850 (1961).
10. Iwahara, J., Schwieters, C. D. & Clore, G. M. Ensemble approach for NMR structure refinement against 1H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule. *J. Am. Chem. Soc.* **126**, 5879–5896 (2004).
11. Iwahara, J., Schwieters, C. D. & Clore, G. M. Characterization of nonspecific protein-DNA interactions by 1H paramagnetic relaxation enhancement. *J. Am. Chem. Soc.* **126**, 12800–12808 (2004).
12. Iwahara, J., Anderson, D. E., Murphy, E. C. & Clore, G. M. EDTA-derivatized deoxythymidine as a tool for rapid determination of protein binding polarity to DNA by intermolecular paramagnetic relaxation enhancement. *J. Am. Chem. Soc.* **125**, 6634–6635 (2003).
13. Gehring, W. J. et al. Homeodomain-DNA recognition. *Cell* **78**, 211–223 (1994).
14. Billeter, M. et al. Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex. *J. Mol. Biol.* **234**, 1084–1093 (1993).
15. Fraenkel, E. & Pabo, C. O. Comparison of X-ray and NMR structures for the Antennapedia homeodomain-DNA complex. *Nature Struct. Biol.* **5**, 692–697 (1998).
16. Clore, G. M. & Garrett, D. S. R-factor, free R and complete cross-validation for dipolar coupling refinement of NMR structures. *J. Am. Chem. Soc.* **121**, 9008–9012 (1999).
17. Iwahara, J. & Clore, G. M. Direct observation of enhanced translocation of a homeodomain between DNA cognate sites by NMR exchange spectroscopy. *J. Am. Chem. Soc.* **128**, 404–405 (2006).
18. Affolter, M., Percival-Smith, A., Müller, M., Leupin, W. & Gehring, W. J. DNA binding properties of the purified Antennapedia homeodomain. *Proc. Natl Acad. Sci. USA* **87**, 4093–4097 (1990).
19. Catron, K. M., Iler, N. & Abate, C. Nucleotides flanking a conserved TAAT core dictate the DNA binding specificity of three murine homeodomain proteins. *Mol. Cell. Biol.* **13**, 2354–2365 (1993).
20. Misteli, T. Protein dynamics: implications for nuclear architecture and gene expression. *Science* **291**, 843–847 (2001).
21. Lewin, B. *Genes VII* (Oxford Univ. Press, Oxford, 2000).
22. Gillespie, J. R. & Shortle, D. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J. Mol. Biol.* **268**, 158–169 (1997).
23. Dvoretzky, A., Gaponenko, V. & Rosevear, P. R. Derivation of structural restraints using a thiol-reactive chelator. *FEBS Lett.* **528**, 189–192 (2002).
24. Gabdoulline, R. R. & Wade, R. C. Biomolecular diffusional association. *Curr. Opin. Struct. Biol.* **12**, 204–213 (2002).
25. Donaldson, L. W. et al. Structural characterization of proteins with an attached ATCUN motif by paramagnetic relaxation enhancement NMR spectroscopy. *J. Am. Chem. Soc.* **123**, 9843–9847 (2001).
26. Schwieters, C. D., Kuszewski, J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–73 (2003).
27. McConnell, H. M. Reaction rates by nuclear magnetic resonance. *J. Chem. Phys.* **28**, 430–431 (1958).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Intramural Program of the NIH, the NIDDK, and in part by the AIDS Targeted Antiviral Program of the Office of the Director of the NIH (G.M.C.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.M.C. (mariusc@intr.nidk.nih.gov).

An expression of interest

Understanding the regulation of gene expression in stem cells and neurons, and finding ways to manipulate it, are important challenges. **Laura Bonetta** searches out the tools for the job.

Researchers face a difficult task as they try to realize the therapeutic potential of stem cells and neurons. To better understand how to manipulate these cells, they need to monitor the gene-expression patterns, as well as working out how these genes are controlled. But both stem cells and neurons are not easy to maintain in culture, and it is hard to introduce DNA or RNA molecules into them to target specific genes or pathways.

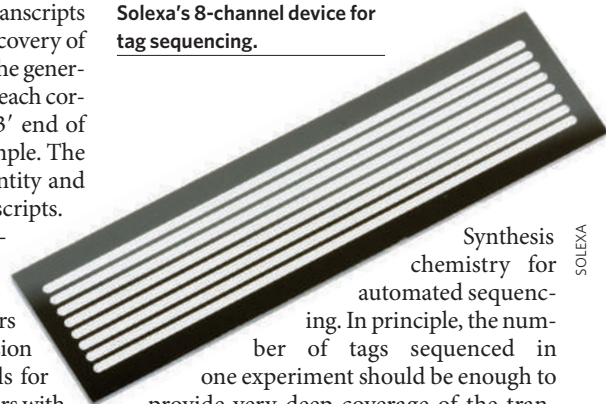
Nonetheless, researchers have been successful in determining the expression of thousands of genes and comparing expression patterns between different cells or cells grown under different conditions. Such work has allowed them to identify, for example, master regulators of stem-cell differentiation or neuronal survival. In addition, a wide variety of tools has been designed specifically for use in stem cells or neurons to control the expression of a gene of interest and study its function.

One useful technique allows scientists to identify all the genes involved in a particular process, such as the migration or differentiation of stem cells. Serial analysis of gene expression (SAGE) is an open platform for monitoring the

expression patterns of thousands of transcripts in one sample and can lead to the discovery of novel genes. The technique relies on the generation of a library of short cDNA 'tags' each corresponding to a sequence near the 3' end of every transcript in a cell or tissue sample. The tags are sequenced to reveal the identity and quantity of the corresponding transcripts. The National Institutes of Health's Cancer Genome Anatomy Project has generated several SAGE human embryonic stem-cell libraries and offers web-based tools to analyse the expression of genes in these libraries (see 'Tools for expression analysis'). To aid researchers with the construction of SAGE libraries, Invitrogen of Carlsbad, California, sells the I-SAGE and I-SAGE Long kits.

A single SAGE experiment generates about 50,000 tags. But in mid-June Solexa in Hayward, California, will launch a sequence-based expression-analysis platform that can analyse more than one million cDNA tags. The Genome Analysis System uses a tag amplification step on the surface of a glass flow cell and features Solexa's Sequencing-by-

Solexa's 8-channel device for tag sequencing.



Synthesis chemistry for automated sequencing. In principle, the number of tags sequenced in one experiment should be enough to provide very deep coverage of the transcripts expressed in a human cell and so should capture those expressed at very low levels, according to Solexa.

Methods such as SAGE and the ever-popular microarrays look at the complement of transcripts isolated from a population of cells. Using a different method, based on gene-trapping technology, scientists at the Salk Institute for Biological Studies in La Jolla, California, have achieved real-time monitoring of gene expression in individual, living mouse neu-

SOLEXA

TOOLS FOR EXPRESSION ANALYSIS

The Mammalian Gene Collection at the National Institutes of Health (NIH) in Bethesda, Maryland, provides full-length, open-reading-frame cDNA clones for human, mouse and rat genes. As part of this effort, NIH-funded scientists have created several cDNA libraries from human embryonic stem-cell lines, which can be studied using tools on the Cancer Genome Anatomy Project website.

- **The Gene Library Summarizer** finds all the genes expressed in a single cDNA library or group of cDNA libraries from different tissues, including stem cells. It then classifies the genes as unique or non-unique, and known or unknown.
- **The cDNA xProfiler** compares gene-expression patterns between two pools of libraries.
- **The Digital Gene Expression Displayer** compares the 'degree' of



The National Institutes of Health's websites offer tools for expression analysis.

presence of a gene in pool A with its degree of presence in pool B. This comparison is reduced to two numbers: the sequence odds ratio and measure of significance.

The Cancer Genome Anatomy Project has also generated several long and one short SAGE (serial analysis of gene expression)

libraries from human embryonic stem cells. The project's website provides tools to analyse the expression of genes represented in these libraries.

- **The SAGE Anatomic Viewer** displays the relative expression of a given gene in different cells. After selecting a gene, the user can

compare the expression in the pool of SAGE embryonic stem-cell libraries to all other major cell lines derived from tissues in the body, both normal and cancerous, or to a mixture of cell lines and tissues. The viewer also provides access to the **Ludwig Transcript Viewer**, which depicts the transcript that contains the tag and three other theoretical tags downstream of the 3' end; and to the **Digital Northern**, which provides the expression of a gene in each library it is found in.

- **The SAGE Digital Gene Expression Displayer** distinguishes significant differences in gene-expression profiles between two pools of SAGE libraries.
- **The SAGE Absolute Level Lister** lists all SAGE libraries and links to the distribution of transcript expression levels in any given library.

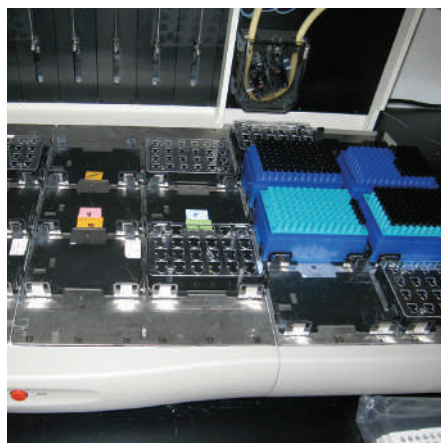
L.B.

ronal stem cells. "When you work with a therapeutically relevant cell system you want to keep it as close to its natural state as possible," says Carolee Barlow, who helped devise the system. Barlow, who is now at Brain Cells in San Diego, California, and her colleagues created a library of stem cells, each with a single retrovirus randomly integrated in its genome. The virus was often integrated within or near a 'trapped' gene, and carried with it a reporter gene that gave off a fluorescent signal when expressed. By detecting fluorescence, the researchers could correlate the expression of a trapped gene with a specific phenotype.

"By PCR or microarray you are looking at genes associated with a specific phenotype," says Barlow. "But by analysing expression in real time you can identify the genes actually driving those phenotypic changes."

Narrowing the search

Large-scale gene-expression screens usually result in a subset of genes that warrant further analysis, with the polymerase chain reaction (PCR) being the usual choice for the first round of follow-up studies. In particular, real-time PCR allows products to be detected as they are being made, which provides a quantitative measurement of expression levels. With some platforms, it is also possible to look at four or five gene targets in a single reaction tube, increasing throughput and lowering costs. For example, QIAGEN in Venlo, the Netherlands, sells a proprietary PCR buffer solution that increases the specificity of each



Liquid handling for Gene Express's Standardized Expression Measurement Center technology.

primer for the respective target sequence, even when several different primer sets are present in a single reaction mix. "It does not matter what primer-probe combinations you use; our product allows for specificity," says Kenneth Dwyer, marketing manager at QIAGEN.

Seegene in Seoul, South Korea, has developed a primer based on dual specific oligonucleotides that allows the length of the primer sequence to be longer than traditional primers in real-time PCR, thereby increasing specificity. "You can use more than five sets of primers in one reaction tube and never have any problems," says Seegene's founder and chief executive, Jong-Yoon Chun. "There is no need for optimization." The product, called

GeneXP, is sold in kits for studying the expression of specific gene families.

For research that may lead to drug development, Gene Express in Toledo, Ohio, markets a technology known as *StaRT*-PCR. Although similar in principle to many other quantitative PCR assays, "the key differentiation is the ability to have an internal standard," according to the company's chief executive Gerald Vardzel. The technology relies on a standardized mix of competitive cDNAs included in all the reaction mixtures, which allows numerical values to be assigned to gene-expression levels and for comparisons to be made across the drug-development pipeline. Gene Express provides the technology as a service through its Standardized Expression Measurement Center.

SuperArray Bioscience in Frederick, Maryland, sells 96-well plates that include real-time PCR primer sets for different panels of pathway- or disease-focused genes. "Even people in the field may not know all the genes related to a particular process. We have done the work for them in identifying the genes of interest," says Sean Yu, vice-president of operations.

The company also sells cDNA and oligonucleotide microarrays for specific sets of genes, including one containing cell-type specific markers for human embryonic stem cells and another that carries representative markers for some of the neural phenotypes. The latter can distinguish between dopaminergic neurons, glial cells and pluripotent stem cells by their gene-expression profiles in a concentration-dependent manner. "The goal was to

GENE EXPRESS

HIGH-CONTENT SCREENING

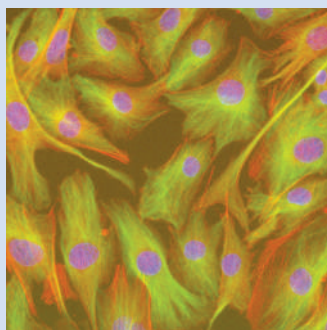
Analysis of stem-cell growth and differentiation requires experiments that manipulate a large number of variables. Typically these involve looking at cell morphology and proliferation, as well as the expression patterns of specific genes or proteins. But Cellomics in Pittsburgh, Pennsylvania, has developed a method for high-content screening that is based on automated image acquisition, processing and analysis.

"The ability to look at many things at once saves time and takes away the subjectivity," says Judy Masucci, marketing director at Cellomics. "A user sets up the criteria for what is a 'hit' and goes home. The next day the machine has cranked through all the plates and the analysis is done."

The ArrayScan VTI HCS Reader takes images of cells on slides or in microplates using multiple fluorescence channels. For example, human embryonic stem-

cell colonies have been cultured on a feeder layer of embryonic mouse fibroblasts in 6-well microplates. The cells were fixed and processed so that their nuclear DNA was labelled. In addition, two other markers — the Oct-4 transcription factor (a marker of pluripotency) and phospho-histone H3 (a mitotic cell marker) — were labelled using standard immunofluorescence procedures. Analysis of images combining these parameters quantified differences in pluripotency among the various cells, the company says.

According to Masucci, what distinguishes the Cellomics idea from similar platforms is the breadth of image-analysis modules and informatics software. Different analysis modules called BioApplications are used depending on the measurements required. "We have about 20 different BioApplications and, depending on which one you're



This image of fluorescently labelled cells (left) was captured using Cellomics's ArrayScan VTI HCS Reader (right).



using, you can measure tens to hundreds of different parameters," Masucci says.

Some BioApplications are very general, looking at a wide variety of parameters, whereas others are more streamlined, such as the one for analysing extended neurite outgrowth. The instrument comes in three different platforms. The ArrayScan VTI HCS Reader is an automated fluorescence system

primarily for fixed cells. It can analyse up to 26,000 wells a day and offers flexibility in terms of imaging applications. The KineticScan HCS Reader is used for live cell biology. The most recent addition to the family, cellWoRx, is cheaper but less flexible. In addition to instruments and bioinformatics, Cellomics offers a database for storing and managing image data.

L.B.

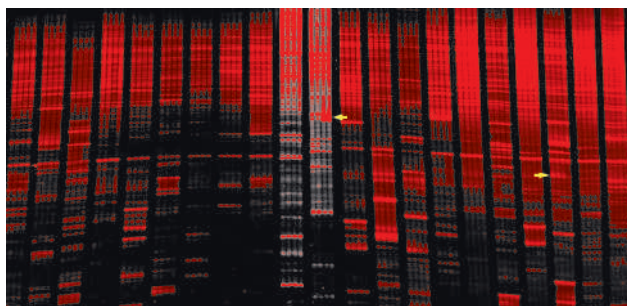
CELLOMICS

provide a practical test for stem-cell differentiation,” says Yu.

Along the same lines, OriGene in Rockville, Maryland, offers TissueScan Real-Time PCR panels to study gene expression in human and mouse tissues. Each array contains PCR-ready cDNAs normalized with beta-actin. For example, the Human Brain TissueScan array contains first-strand DNA from 24 tissues in the human brain. The company markets TissueScan Real-Time PCR disease panels for gene-expression analyses across many stages of disease. “It is prohibitively expensive to get hold of high-quality disease tissues,” says Rich Hamer, vice-president for business development at OriGene.

Comparing expression

For comparing genes expressed by two different cells, for example an embryonic stem cell and one that has begun to differentiate, the usual method is differential display. Developed in 1992, the technology works by systematically amplifying portions of mRNAs from two or more samples and resolving the amplified products by denaturing polyacrylamide gel electrophoresis (PAGE). This allows sequence information to be recovered and corresponding cDNAs to be isolated for further molecular and functional characterizations. “It is especially good for samples with limited RNA or species that have not been sequenced,” says Jonathan Meade, product manager for



GenHunter differential display showing changes in gene regulation.

GenHunter in Nashville, Tennessee. GenHunter sells two kits for differential display: the RNAimage Kit uses radioactive detection, and the RNAspectra Kit uses fluorescence.

Seegene's GeneFishing DEG (differentially expressed gene) discovery kit also relies on amplification with randomly designed primers. But this method uses annealing control primers, which are longer than primers used in differential display and so provide greater annealing specificity during PCR. The resulting PCR products are sufficiently long to be detected on an agarose gel. “The big advantage is that you can use agarose gels, which are easier to prepare than PAGE,” says Chun.

In most cases, microarrays and quantitative PCR technologies cannot assign a gene-expression status to a specific cell type. A better choice for this purpose is fluorescent *in situ* hybridization (FISH). This uses fluorescently labelled DNA oligonucleotide (or RNA) probes to visualize the expression of genes in a single cell with

a microscope. Gene expression measured by FISH can be combined with other visual properties of the cell in a high-throughput, automated fashion (see ‘High-content screening’).

Aureon Laboratories in Yonkers, New York, has developed a method called peT-FISH (paraffin embedded tissue FISH) for detecting signals from nascent RNA molecules localized at the transcription site of genes being expressed. “We are looking at the presence of nascent RNA that

correlates with early events of gene regulation in response to stimuli,” says Paola Capodiceci, one of Aureon's scientists. “The transcript is still in the nucleus and we don't know if it will become mRNA or protein. So we can see what happens at the beginning of the expression chain.” So far, the method has detected up to five genes in a single cell. One of the advantages of this technology is that tissue is not destroyed. “We preserve the tissue morphology,” says Capodiceci.

Homing in on individual genes

To study the function of a gene, researchers typically alter its expression, either over-expressing it or knocking it out. A popular way to turn down the expression of specific genes is through RNA interference. This uses synthetic double-stranded RNA oligonucleotides known as small interfering RNAs (siRNAs), short-hairpin RNAs (shRNAs), which are expressed from a polIII promoter on a plasmid

A MICRO PERSPECTIVE ON STEM CELLS

MicroRNAs (miRNAs) are single-stranded RNA molecules roughly 22 nucleotides long that are generated from endogenous hairpin-shaped transcripts encoded in the genomes of humans, animals, plants and viruses. In mammalian cells, the miRNA molecule is incorporated into a protein complex that usually inhibits translation of mRNAs containing sequences partially complementary to it. A number of processes are regulated by miRNAs, including apoptosis, differentiation and the maintenance of stem-cell properties.

To investigate the function of miRNAs in stem cells, scientists at Applied Biosystems in Foster City, California, have developed a real-time PCR-based method to monitor the expression of 216 miRNA molecules in a single mouse embryonic stem cell. “If you study gene expression in stem

cells, you don't want to look at a population of cells that may include stem cells and differentiated cells,” says Kaiqin Lao, a scientist at the company. The technology Lao developed in collaboration with Azim Surani at the University of Cambridge, UK, involves hand-picking a single cell and, after lysis, performing real-time PCR. “The way we designed the assay is such that you only

amplify miRNAs, so you don't need purification,” says Lao. Although kits for PCR analysis in single cells are not commercially available, Applied Biosystems has launched a line of TaqMan miRNA assays for quantifying miRNAs by real-time PCR.

Several other companies also offer tools for miRNAs. Ambion in Austin, Texas, has a new line of products “to upregulate and

downregulate miRNA expression”, explains Ellen Prediger, senior manager of technical publications at the firm. “One can then assess the role specific miRNAs play in biological processes, including disease.” Ambion sells sets of Pre-miR miRNA precursors, which are double-stranded, chemically modified nucleic acids designed to mimic endogenous mature miRNA molecules. It also makes anti-miR miRNA inhibitors — nucleic acids designed to specifically target miRNA molecules in cells.

Invitrogen of Carlsbad in California offers a set of cloned miRNAs in plasmid vectors. Unlike short-hairpin RNAs, miRNAs are expressed from polIII promoters, and can easily be transferred to lentiviral vectors or vectors with inducible promoters. “We are launching a number of collections of miRNAs for druggable targets,” says Peter Welch, the company's research director.



Ambion offers products to aid the investigation of microRNAs.

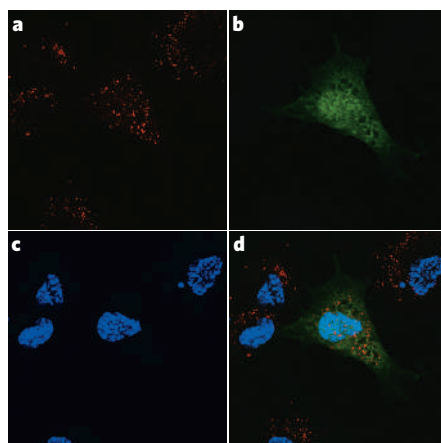
vector, or microRNAs (see 'A micro perspective on stem cells'). But slow-growing stem cells and neuronal cells are difficult to transfect with these molecules.

One way to introduce plasmids or siRNAs into cells is to use lipid-based transfection products. These work by forming a complex with DNA or RNA that interacts with the cell membrane. Invitrogen's Lipofectamine 2000 works with many mammalian cells, and can introduce an siRNA corresponding to the transcription factor Oct-4 in stem cells. "We got about 80% delivery," says Peter Welch, director for research and development.

Several companies have developed transfection reagents to introduce plasmids and siRNAs specifically into cells that are hard to transfect. Mirus Bio in Madison, Wisconsin, has the *TransIT* LT reagent for neuronal cell lines, which is a combination of an endogenous cellular protein, histone H1 and lipoamine. "The histone H1 is the primary cationic carrier and reduces the amount of lipid needed, reducing toxicity," says James Hagstrom, vice-president for scientific operations at Mirus.

This year, Panomics in Fremont, California, will launch a peptide-based reagent called DeliverX, which it claims will have high efficiency and low toxicity. "It has been validated with one primary cell and this summer more primary and suspension cells will be added," says Ian Ley, the company's vice-president for marketing.

When transfection agents fail, electroporation is the brute-force approach for delivering



DNA delivery directly into the nucleus visualized with different fluorescent labels.

genetic material inside cells. The nucleofector technology marketed by Amaxa Biosystems of Cologne, Germany, is aimed at cell lines that are difficult to transfect. A series of gentle pulses allows transfected DNA directly to enter the nucleus. "Other non-viral transfection methods rely on cell division for the transfer of DNA into the nucleus. Nucleofection provides the ability to transfect even non-dividing cells such as neurons and resting blood cells," says spokeswoman Kimberly Stevenson. Depending on the cell type, a researcher would choose one of many proprietary nucleofector solutions and program the machine with the appropriate parameters. For experiments that require stably transfected

cells or *in vivo* work, viral vectors are the obvious choice (see 'Hitching a ride on a virus').

Regardless of the method of gene expression used, some steps in the process will probably require growing cells in culture, using formulated liquid media supplemented with growth factors and other substances that promote cellular replication and govern differentiation. The conditions and the reagents used may affect the gene-expression programmes of a cell, a concern that is particularly relevant to stem-cell research. "For the stem-cell field to move forward it is necessary to have standardized reagents and assays to allow results to be compared," says Sharon Louis, senior scientist at StemCell Technologies in Vancouver, Canada. "Where we see ourselves is providing standardized tissue-culture reagents." The media reagents provided by the company come with detailed protocols for either maintaining cells in an undifferentiated pluripotent state or for inducing differentiation towards a specific lineage.

Each in their own way, stem cells and neuronal cells present problems for the scientists who want to manipulate them. But many scientists and companies have risen to this challenge by developing increasingly sophisticated tools and technologies. With this tool kit in hand, studying gene expression in stem and neuronal cells is a realistic goal that has been embraced by researchers with an enthusiasm matching the promises of their field.

Laura Bonetta is a freelance writer based in the Washington DC area.

HITCHING A RIDE ON A VIRUS

Nucleic acids can be introduced into cells by allowing them to hitch a ride on a virus. Typically this involves engineering a viral vector to carry a gene of interest as well as a selectable marker or reporter gene. The vector is then introduced into a packaging cell line to produce recombinant viral particles, which are used to infect the cells of interest.

Many products on the market provide complete kits that include vector and reporter plasmids, packaging cells, PCR primers, as well as all the necessary buffers and solutions. When working with non-dividing or slow-growing cells, the best choices are systems that are based on lentiviruses or adeno-associated viruses (AAVs).

Lentiviruses, a subclass of retroviruses, can infect cells that are not dividing and are able to integrate DNA into the host cell. The ViraPower lentiviral expression system made by



Didier Trono has helped to devise a novel lentiviral vector.

Invitrogen of Carlsbad, California, delivers and actively transports a gene of interest to the nucleus of a broad range of cells, including stem cells and neuronal cells.

AAV vectors can also introduce genes into these cells but there is a downside: the virus typically needs to be co-infected with adenovirus to produce recombinant AAV virions in packaging cells. One possible way around this is to use the AAV helper-free system sold by Stratagene of La Jolla, California. This eliminates the need for the helper virus, greatly simplifying the whole process.

Scientists at the Swiss Federal Institute of Technology in Lausanne have developed a lentiviral vector that they say allows regulated expression of genes or short-hairpin RNAs (shRNAs) in any cell type or tissue either *in vitro* or *in vivo* using ubiquitous or tissue-specific promoters. The technique takes advantage of the repression activity of the Krüppel-associated box (KRAB) domain. This domain is found in many vertebrate

transcriptional regulators and, when it is tethered to DNA, it induces a local heterochromatin state that represses transcription.

By coupling the KRAB domain to a tetracycline operator sequence on the viral vector, expression of any gene carried by the virus is blocked. The block is readily removed by adding an antibiotic. This allows gene expression to be switched on and off at will. "The main advantage of this vector is the ability to control polIII promoters, such as those that regulate shRNA expression, and to control polIII promoters, which can serve to express transgenes or miRNAs. It provides a means to do drug-inducible shRNA knockdown," says Didier Trono, dean of the School of Life Sciences at the Lausanne institute, who led the team that came up with the technique. "It has a range of applications far greater than existing systems."

L.B.

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL
PCR reagents and equipment			
ABgene	PCR reagents, quantitative PCR, microplates	Epsom, UK	www.abgene.com
AlleLogic Biosciences	Real-time PCR, AllGlo technology	Hayward, California	www.allelogic.com
Althea Technologies	eXpress Profiling for high-throughput gene expression analysis	San Diego, California	www.altheatech.com
Ambion	Reverse transcription kits and reagents, enzymes and other products for PCR and RT-PCR	Austin, Texas	www.ambion.com
Applied Biosystems	Real-time PCR instruments, TaqMan primers and probes for real-time PCR, kits for microRNA PCR	Foster City, California	www.appliedbiosystems.com ●
Beckman Coulter	GenomeLab GeXP genetic-analysis system for highly multiplexed PCR	Fullerton, California	www.beckmancoulter.com
Bio-Rad	Enzymes and kits for PCR, thermal cyclers	Hercules, California	www.bio-rad.com ●
Cepheid	Equipment for DNA analysis and PCR including SmartCycler system for real-time PCR	Sunnyvale, California	www.cepheid.com
Chemicon	PCR kits	Temecula, California	www.chemicon.com
Clontech	Enzyme mixes for PCR, custom PCR service, PCR-Select cDNA Subtraction Kit	Mountain View, California	www.clontech.com
Corbett Life Science	Rotor-Gene 3000 and 6000 for real-time PCR	Sydney, Australia	www.corbettlifescience.com
EPICENTRE Biotechnologies	Enzymes for PCR and RT-PCR, DNA and RNA purification	Madison, Wisconsin	www.epibio.com
Epoch Biosciences	Custom oligonucleotides, MGB Eclipse probe systems	Bothell, Washington	www.epochbio.com
Exiqon	Locked nucleic acid oligonucleotides, ProbelLibrary gene-expression kits for real-time PCR	Vedbaek, Denmark	www.exiqon.com ●
Finnzymes	Phusion and DyNAzyme DNA polymerases for PCR, DyNAmo qPCR Kits for real-time PCR	Espoo, Finland	www.finnzymes.fi
GE Healthcare	PuReTaq Ready-To-Go PCR beads, PCR master mixes, oligonucleotide synthesis, PCR clean-up reagents	Little Chalfont, UK	www.gehealthcare.com ●
Gene Express	StaRT-PCR (standardized reverse transcriptase-polymerase chain reaction) technology	Toledo, Ohio	www.geneexpressinc.com
GenScript	PCR cloning, subcloning and mutagenesis service, TissueDirect multiplex PCR system	Piscataway, New Jersey	www.genscript.com
Idaho Technology	RapidCycler instruments for real-time PCR	Lake City, Utah	www.idahotech.com
Integrated DNA Technologies	Custom oligonucleotides including locked nucleic acids	Coralville, Iowa	www.idtdna.com
Invitrogen	Enzymes and kits for PCR, cloning, genomics, proteomics, genotyping kits, antibodies	Carlsbad, California	www.invitrogen.com
Lucigen	EconoTaq DNA polymerase	Middleton, Wisconsin	www.lucigen.com
Maxim Biotech	Kits for real-time and RT-PCR, custom primer design and optimization, RT-cDNA synthesis	Rockville, Maryland	www.maximbio.com
Megabase Research Products	PCRJet pressurized gas thermocycler for high-speed amplification of DNA	Lincoln, Nebraska	www.pcrjet.com
MIDSCI	Laboratory supplies including PCR instruments	St Louis, Missouri	www.midsci.com
MWG Biotech	Custom oligonucleotides, probes for quantitative PCR, microarrays	Ebersberg, Germany	www.mwg-biotech.com
New England Biolabs	Various qPCR and RT-PCR reagents, Taq and Deep VentR DNA polymerases	Ipswich, Massachusetts	www.neb.com
Novagen	Thermostable DNA polymerases, BloodDirect PCR buffer kits, RT-PCR kits, primers, PCR clean-up	Madison, Wisconsin	www.emdbiosciences.com
OriGene	TissueScan real-time PCR panels and TissueScan real-time PCR disease panels	Rockville, Maryland	www.origene.com
Panomics	QuantiGene reagent system for gene-expression analysis, delivery kits for siRNA	Fremont, California	www.panomics.com
Promega	PCR kits, pGEM-T and pGEM-T Easy Vector systems for cloning PCR products, PCR clean-up reagents	Madison, Wisconsin	www.promega.com
Qbiogene	DNA polymerases and master mixes for real-time and RT-PCR, &GO master mixes, Taq CORE kits	Irvine, California	www.qbiogene.com
QIAGEN	Kits for real-time and RT-PCR, multiplex PCR kit, reagents and kits	Valencia, California	www1.qiagen.com
R&D Systems	Primer pairs for RT-PCR, Quantikine mRNA quantitation kits	Minneapolis, Minnesota	www.rndsystems.com
Roche Applied Science	Enzymes and kits for PCR, LightCycler thermal cycler for real-time PCR	Lewes, UK	www.roche-applied-science.com ●
Seegene	GeneFishing kit for identifying differentially expressed genes, and GeneXP multiplex RT-PCR system	Seoul, South Korea	www.seegene.com
Sigma-Aldrich	PCR buffers and enzymes, JumpStart Taq products, AccuTaq LA and KlenTaq LA enzymes	St Louis, Missouri	www.sigmaaldrich.com ●
Stratagene	Enzymes and kits for PCR, Mx3000P and Mx4000 real-time PCR systems	La Jolla, California	www.stratagene.com ●
SuperArray Bioscience	PCR arrays for analysing the expression of a focused panel of genes	Frederick, Maryland	www.superarray.com ●
Takara Bio	Buffers and enzymes for long and accurate PCR, kits for real-time and RT-PCR, PCR thermal cycler dice	Shiga, Japan	www.takara-bio.com ●
Techn	Thermal cyclers	Burlington, New Jersey	www.techn.com
Thermo Electron	PCR Sprint and MBS satellite thermal cyclers, custom peptide and oligonucleotide synthesis	Ashford, Middlesex	www.thermo.com ●
Wako Chemicals	Duplex-specific Direct Digestion (DsDD) subtraction kit for identifying differentially expressed genes	Richmond, Virginia	www.wakousa.com

Reagents for RNA interference

Ambion	siRNAs to human, mouse and rat genes, siRNA libraries, custom siRNAs, siRNA expression vectors, kits for preparing siRNAs, kits for miRNA detection and isolation	Austin, Texas	www.ambion.com
BD Biosciences	Culture media, radioassay kits, FACS range of flow cytometers, RNAi kits	San Jose, California	www.bd.com
Benitec	Gene-silencing RNAi technology	Mountain View, California	www.benitec.com
Bioneer	siRNA and oligo design service	Daejeon, South Korea	eng.bioneer.com
Cenix BioScience	Genome-based high-throughput applications of RNA interference	Dresden, Germany	www.cenix-bioscience.com
Dharmacon	RNA technologies, siRNA kits and arrays, custom RNAs, large-scale RNA synthesis	Lafayette, Colorado	www.dharmacon.com
Eurogentec	Custom oligos and siRNAs, kits for qPCR and RTqPCR	Seraing, Belgium	www.eurogentec.be
Exiqon	miRNA microarrays, microRNA profiling service using miRCURY LNA array	Vedbaek, Denmark	www.exiqon.com ●
Imgenex	Plasmid-based siRNA kits; polyclonal and monoclonal antibodies; ELISA kits	San Diego, California	www.imgenex.com
Invitrogen	siRNA and miRNA products, including BLOCK-iT RNAi vectors, Stealth siRNA collections for specific gene families, viraPower lentiviral expression system, NCode miRNA labelling system	Carlsbad, California	www.invitrogen.com
LC Sciences	microRNA detection microarray service	Houston, Texas	www.lcsiences.com
Molecula	Custom antisense oligos, siRNAs, pDsIPHER shRNA vector, transfection reagents	Columbia, Maryland	www.molecula.com
OligoEngine	Custom oligos for RNAi, DNA methylation, antisense and genome arrays	Seattle, Washington	www.oligoengine.com
Open Biosystems	MicroRNA-adapted shRNA libraries, non-microRNA-adapted shRNA libraries, RNAi transfection kits	Huntsville, Alabama	www.openbiosystems.com
OriGene	shRNA expression plasmids and retroviral vectors, matching cDNA clones and antibodies	Rockville, Maryland	www.origene.com
Proligo	Custom DNA and RNA oligos for RNAi, LNA oligos, bi-labelled fluorescent probes for PCR	Boulder, Colorado	www.proligo.com
QIAGEN	Genomewide siRNAs, siRNA sets for specific gene families, HiPerFect transfection reagent	Venlo, the Netherlands	www1.qiagen.com
RNA-TEC	Custom oligonucleotide synthesis, RNAi for <i>in vivo</i> applications	Leuven, Belgium	www.rna-tec.com
Rosetta Genomics	microRNA discovery, validation, bioinformatics and microarray development	Rehovot, Israel	www.rosettagenomics.com

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL
Spring Bioscience	Kits for generating siRNA	Fremont, California	www.springbio.com
SuperArray Bioscience	SureSilencing shRNA plasmids	Frederick, Maryland	www.superarray.com
System Biosciences	Kit for global amplification of the complete small RNA transcriptome, genome-wide siRNA libraries, lentiviral vectors for siRNA, RNA amplification kits, cloning vectors	Mountain View, California	www.systembio.com
FISH reagents and imaging systems			
Abbott Molecular	<i>In situ</i> hybridization reagents, FISH pretreatment reagents and kits, fluorescence-labelling reagents	Des Plaines, Illinois	www.vysis.com
Applied Precision	Image-analysis software, microarray scanning and analysis systems, high-content cell analysis	Seattle, Washington	www.api.com
Aureon Laboratories	Paraffin-embedded tissue FISH, multicolour fluorescent <i>in situ</i> protein detection platform, MAGIC automated machine vision analysis system	Yonkers, New York	www.aureon.com
Cellomics	High-content screening platforms, image-analysis software, image-storage system	Pittsburgh, Pennsylvania	www.cellomics.com
Exiqon	LNA probes for RNA and DNA <i>in situ</i> hybridization	Vedbaek, Denmark	www.exiqon.com
Genisphere	3DNA FISH probes, fluorescently labelled positive and negative control kits, RNA amplification	Hatfield, Pennsylvania	www.genisphere.com
KPL	Reagents for <i>in situ</i> hybridization and detection of biotinylated DNA in cells and tissues	Gaithersburg, Maryland	www.kpl.com
Molecular Probes	FISH Tag DNA and FISH Tag RNA kits optimized for multiplex fluorescence <i>in situ</i> hybridization	Carlsbad, California	probes.invitrogen.com
MP Diagnostics	FISH reagents and supplies, FISH probes for specific applications	Irvine, California	www.mpdiagnostics.com
Vector Laboratories	Reagents for FISH, immunofluorescence, immunohistochemistry, neuronal tracing, flow cytometry, ELISA	Burlingame, California	www.vectorlabs.com
Stem cell culture and isolation			
American Type Culture Collection	ATCC ELF phosphatase-detection-kit assay, cell-culture media, cell-biology assays	Manassas, Virginia	www.atcc.org
Fisher Scientific	Cell-culture products	Hampton, New Hampshire	www.fishersci.com
Genetix	ClonePix for picking stem-cell clones from plate into microwell plates	New Milton, UK	www.genetix.com
Miltenyi Biotec	Immunomagnetic isolation, phenotypic analysis, cultivation and molecular analysis of cells	Gladbach, Germany	www.miltenyibiotec.com
R&D Systems	Media products for stem and progenitor cell culture, kits for differentiation, expansion and identification, stem-cell antibody panels	Minneapolis, Minnesota	www.rndsystems.com
Serologicals	Products to culture, analyse and differentiate both adult and embryonic stem cells	Norcross, Georgia	www.serologicals.com
Specialty Media	EmbryoMax ES cell media and reagents, kits for gene knockout, genomic libraries, vectors, mouse embryonic stem-cell lines	Phillipsburg, New Jersey	www.specialtymedia.com
StemCell Technologies	Media and cell-separation products, antibodies, tissue-culture reagents and contract services	Vancouver, Canada	www.stemcell.com
Gene delivery systems			
Amata	Nucleofector technology	Cologne, Germany	www.amata.com
Avanti Polar Lipids	Transfection reagents	Alabaster, Alabama	www.avantilipids.com
Bio-Rad	Nucleic-acid delivery systems and reagents, imaging systems and software	Hercules, California	www.bio-rad.com
BioVision	GeneBlocker pGB siRNA expression vectors	Mountain View, California	www.biovision.com
BTX Molecular Delivery Systems	Electroporation and cell-fusion systems	Holliston, Massachusetts	www.btxonline.com
Clontech	Retroviral and adenoviral expression systems, Tet-On and Tet-Off gene-expression systems	Mountain View, California	www.clontech.com
Fermentas	Transfection reagents, oligos, reagents for PCR	Burlington, Canada	www.fermentas.com
GeneChoice	Transfectol transfection reagent	Frederick, Maryland	www.genechoiceinc.com
Genlantis	GenePORTER 2 transfection reagent, NeuroStem neuronal progenitor cells from rodent brain tissue, NeuroFECT transfection reagent for plasmid delivery to primary neuronal cells	San Diego, California	www.genlantis.com
Invitrogen	BLOCK-iT Pol II miR RNAi expression vector kits and BLOCK-iT shRNA vector kits, ViraPower lentiviral expression system, transfection reagents	Carlsbad, California	www.invitrogen.com
InvivoGen	Cloning vectors including vectors for siRNA	San Diego, California	www.invivogen.com
Microbix	Adenovirus kits	Toronto, Canada	www.microbix.com
Targeting Systems	Gaussia luciferase-based expression vectors for studying gene expression	Santee, California	www.targetingsystems.com
Mirus Bio	<i>In vitro</i> transfection reagents, <i>in vivo</i> delivery vectors, DNA and RNA labelling kits	Madison, Wisconsin	www.mirusbio.com
MoBiTec	DNA vector systems	Göttingen, Germany	www.mobitec.de
OZ Biosciences	Transfection reagents for different cell lines and applications, magnofection technology	Marseille, France	www.ozbiosciences.com
Polyplus-transfection	Reagents for <i>in vitro</i> and <i>in vivo</i> delivery of genes, oligonucleotides, siRNA and proteins	Illkirch, France	www.polyplus-transfection.com
Promega	Vectors for cDNA cloning, two-hybrid screening, mammalian expression vectors, RNAi vectors	Madison, Wisconsin	www.promega.com
Stratagene	Viral gene delivery systems for mammalian cells, two-hybrid products, transfection reagents	La Jolla, California	www.stratagene.com
Tritech Research	Electroporation equipment and microinjectors	Los Angeles, California	www.tritechresearch.com
Upstate	Transfection reagents, antibodies, cell-culture reagents	Charlottesville, Virginia	www.upstate.com
VIRxSYS	Genetic therapy development	Gaithersburg, Maryland	www.virxsys.com
Other gene expression analysis			
BioHelix	IsoAmp tHDA kit amplification kit based on the helicase-dependent amplification method	Beverly, Massachusetts	www.biohelix.com
DNAFORM	Full-length cDNA clones, kits for plasmid DNA preparation, PCR products, ORF cloning, clone handling	Tokyo, Japan	www.dnaform.jp
Evrogen	Fluorescent proteins, cDNA library subtraction service for identifying differentially expressed mRNAs	Moscow, Russia	www.evrogen.com
GenHunter	Kits for identifying differentially expressed mRNAs; PCR cleaning, cDNA probe labelling, RNA isolation	Nashville, Tennessee	genhunter.com
Nanosphere	Verigene System for the automated analysis and detection of nucleic acids and proteins	Northbrook, Illinois	www.nanosphere-inc.com
NUGEN Technologies	WT-Ovation RNA amplification system for whole transcriptome gene expression analysis	San Carlos, California	www.nugentechnologies.com
Solexa	Genome analysis system for expression profiling, small RNA discovery and analysis and sequencing	Little Chesterford, UK	www.solexa.com
SeqWright	Sequencing of SAGE libraries, siRNA expression vector construction, Affymetrix microarray service	Houston, Texas	www.seqwright.com
US Genomics	Trilogy 2020 single molecule analyser	Woburn, Massachusetts	www.usgenomics.com

● see advertisement

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

A howl of hope

This year is the 50th anniversary of the landmark poem *Howl* by Allen Ginsberg, whose opening line is "I saw the best minds of my generation destroyed by madness." So it's timely to see how the verse (not including the stuff about sex and drugs, which is, admittedly, about 75% of it) might apply to the next generation of scientists.

In the first part of *Howl*, Ginsberg writes about a generation who "passed through universities with radiant cool eyes". So far, so good. One can interpret this as guarded optimism in the quest for knowledge, perhaps exemplified by today's student pipeline, which is producing an ample number of scientists.

A phrase halfway through the middle section, "ashcans and unobtainable dollars", is a bit foreboding, especially if you consider the flattening of the US National Institutes of Health research budget to be equivalent to unobtainable dollars that could put young researchers' careers into the ashcan.

The next section is even more ominous, as it rages against the machine of "Moloch", a god who requires human sacrifice. In science, this could include a study

section, an institutional review board or a regulatory agency, all of which have the power to halt a research project as they see fit.

But here, too, there's hope. Scientists have built up ways to appease, if not beat, their individual Molochs by banding together under student and postdoc organizations, professional societies, lobbying efforts and mentoring schemes.

Howl chronicles how an individual can be crushed by an institution. It could be read as a warning for scientists to band together, avoid alienation and triumph over obstacles. Perhaps scientists who avoid isolation could rewrite the cautiously triumphant ending as follows:

"Holy the supernatural extra brilliant intelligent funding of the grant."



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

US Head Office, New York
 75 Varick Street, 9th Floor,
 New York, NY 10013-1917
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager/Corporations:
 Peter Bless
 Classified Sales Representatives
 Tel: +1 800 989 7718

**New York/Pennsylvania/
 Latin America:** Andrew Bennie
**Midwest USA/Maryland/
 NIH:** Wade Tucker
East USA/Canada:
 Janine Taormina

San Francisco Office
Classified Sales Representative:
 Michaela Bjorkman
 225 Bush Street, Suite 1453
 San Francisco, CA 94104
 Tel: +1 415 781 3803
 Fax: +1 415 781 3805
 e-mail: m.bjorkman@naturesf.com

European Head Office, London
 The Macmillan Building,
 4 Crinan Street,
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

European Sales Manager:
 Andy Douglas (4975)

Advertising Production Manager: Billie Franklin
 To send materials use London address above.
 Tel: +44 (0) 20 7843 4814
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
 Patrick Phelan
 e-mail: p.phelan@nature.com

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

MOVERS

Zachary Fisk, distinguished professor of physics, University of California, Irvine



2003–2006: Professor of physics, University of California, Davis

2002–2003: Bernd T. Matthias visiting scholar, Los Alamos National Laboratory, New Mexico

1994–2003: Professor of physics, Florida State University, Tallahassee, Florida

Zachary Fisk has followed the footsteps of some formidable physicists — and then blazed his own trail. Not only was his father president of Bell Labs in New Jersey, Fisk worked at Los Alamos National Laboratory alongside some of the Manhattan Project crowd while taking a break from his undergraduate study at Harvard University. At Los Alamos his interest in synthesizing chemicals created a bond with his mentor, Bernd Matthias. Fisk helped set up Matthias's lab at the University of California, San Diego, and later, as Matthias's graduate student, Fisk made huge numbers of intermetallic compounds to explore their superconducting and magnetic properties.

After Matthias died, Fisk was offered a job at Los Alamos. He worked out a 'poor man's' technique for growing crystals from molten metals as they cool. The soon-to-become Los Alamos director Siegfried Hecker wanted to study radioactive metals for bomb research, so Fisk grew crystals to help him. His work took on an exploratory nature: "I'm more of a hunter-gatherer," he says of his interest in finding systems where physics plays out simply. Fisk sought to legitimize the military work to the community by promoting its societal applications — in the process establishing Los Alamos as a powerhouse in condensed matter physics.

Later, Fisk followed the National High Magnetic Field lab's move to Florida State University at Tallahassee to take advantage of its magnetic probes. These are useful for experiments with heavy-fermion materials whose properties lie at the boundary of magnetism — his forte.

Indeed, Fisk has met with success working at the boundaries of science. "Bridging the gap between what solid-state chemists and condensed matter physicists are interested in was not easy," he says. "But it was where the interesting science was happening."

Family considerations brought him to the University of California, Davis, two years ago. Eager to get back to the San Diego region, Fisk recently joined the faculty at University of California, Irvine, where he plans to continue his lifelong work on heavy-fermion intermetallics and their superconductive roots. Fisk has never forgotten Matthias's advice to comb the literature for data, not conclusions. Matthias contended that people are much better at measuring things than understanding their significance. Fisk agrees. "Developing a nose for when something doesn't add up allows you to go to new places," he says. **Virginia Gewin**

MENTORS & PROTÉGÉS

Thank you, Professor MacDonald

Good mentors are hard to find. As graduate students in Clint MacDonald's laboratory, we recognize the positive effect that expert mentoring can have on a student's career. In the past Clint has won the graduate school's Teacher of the Year award here at Texas Tech University Health Sciences Center, showing that others also know he is a great teacher. He has been described as humble, patient and approachable. However, it is his sincerity that makes him a great mentor.

Professor MacDonald constantly works with us when he is busy. He spent weekends helping Ebtesam Attaya perfect her dissertation defence presentation. He spent an hour introducing Ganesh Shankarling to the lab despite a looming grant-application deadline. He literally has an open-door policy and despite being one of the most productive scientists in our department (he was awarded an Independent Scientist Award from the National Institutes of Health in addition to his ample laboratory funding), he takes time for all students here.

MacDonald is not just available to talk science. All of us have benefited from his countless notes on how to make a better presentation, poster or paper. Andrew Hockert said that MacDonald's long hours and helpful advice allowed him to give his best

seminar ever. MacDonald gives abundant notes on every abstract that leaves the lab, always teaching us how to produce stimulating documents and presentations. His expertise in this area is exemplified in his students: we have won our institution's research-poster competition for the past three years.

However, MacDonald's mentoring is best illustrated in how he develops our careers. He encourages most of us to take extra classes to give us the skills we want to take to our next job, even when such classes require time away from the lab. He also brings us to meetings, introduces us to important researchers and speaks exceedingly highly of each of us during his talks. Even after we leave the lab, he continually writes letters of recommendation, offers career advice and supports us however he can.

MacDonald is a great mentor because he sincerely wants the best for each of us. We are honoured to be able to congratulate and thank Clint for his years of mentoring. We also hope that others too will consider mentoring a key part of their duties as a scientist. ■

Wyatt McMahon, Ebtesam Attaya, Toni Denison, Andrew Hockert and Ganesh Shankarling are graduate students in Clint MacDonald's laboratory at Texas Tech University Health Sciences Center, Lubbock.

GRADUATE JOURNAL

The postdoc menu

Some people take one look at a menu and know exactly what they want. But at every table there's one person who looks at the menu and freezes. They make lists of possibilities, including back-ups in case a dish has run out. They come up with formulas that take into account their affinity for each ingredient and when they last ate a similar dish. Invariably they have a few food allergies too.

When you're a grad student, picking a postdoc is like choosing dinner at a particularly huge restaurant. The menu is daunting, and factoring in each ingredient just makes it worse. Should you try the new field you keep reading about (read: Chilean sea bass) or one you've tried before (read: salmon)? Should you choose the lab that addresses interesting questions with novel techniques (read: *sous vide*) or familiar favourites (read: roast)? Is now the time to see whether that allergy to cnidarians was just a phase?

Choosing a postdoc has one notable advantage over restaurant dilemmas: you can apply to more than one place — like taking a taste of all the dishes that intrigue you. Graduate school has made me better at making informed decisions, but deciding on a postdoc isn't going to come down to a magic formula. Still, having sent out application letters for positions this week, I will interact with each lab before committing to my future scientific sustenance. ■

Milan de Vries is a molecular-biology graduate student at the Massachusetts Institute of Technology.

Ringin' up baby

Order a little sister, then order her about.

Ellen Klages

Nanny says that I am spoiled. It comes from being an only child, and not having to share holidays or cakes and always getting to sit by the window. If I had a little brother or sister, I would learn responsibility. More work for her, she sighs, but she is only thinking of my character. Thinking about me is Nanny's job.

Of course, Mother is far too busy to have a baby right now, what with the Henderson case and all. (When I have supper with her, on Wednesdays, she talks about nothing but the Henderson case.) So Nanny has arranged for a nice lady to plant Mother's egg and do all the messy parts, then give the baby to us when it's done.

"What would you like," Nanny asks me over cocoa. "A brother or a sister?"

I have to think for a moment, but only a little, because a brother would be a pest and get into my best things, like Courtney Taylor's brother Robby, who programmed her mobile phone to ring with a nasty farting sound. A sister is someone I can be the boss of.

"A sister, please," I say in my sweet voice. Nanny loves my sweet voice.

Nanny touches a box on the wall screen, and it glows bright pink.

"Birthday?" she asks, her finger not quite touching the screen, but ready.

My birthday is in June. "October," I say after a minute, because I've had to count in my head, so her party won't get in the way of Christmas, either.

"Excellent," says Nanny. "We can place our order today." She taps her finger on the screen. That box glows red.

"What else can we pick?" There are a lot of boxes. I finish my cocoa and stand right next to Nanny, who smells like Vermont. A nice cool green smell.

She begins to read to me, scrolling slowly down.

"Hair colour?"

"Brown." Mine is honey blond.

"Eyes?"

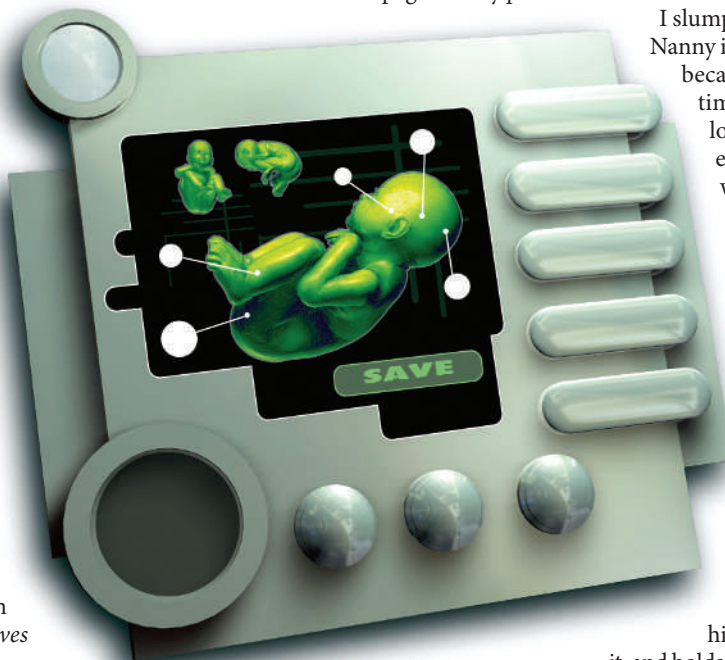
Mine are blue, so brown again. "Intelligence?"

I have to think about that. I don't want a sister who's *stupid*, but if she's smarter than me, she will be difficult to boss.

"Above average," Nanny decides. "Good at maths?"

Hmm. I'm in second grade, and we're doing the times tables. That could be useful. But it probably isn't something she'll be able to do right away.

So I shrug, which is a mistake, because Nanny is very strict about manners and posture and I have to listen to a lecture before she will tap the bottom of the screen and scroll to the next page of baby parts.



This page is less interesting because the words are very long and I don't know what they mean. Bioimmunity. Cholesterol. Neuromuscular. I stare at the screen with my eyes very wide so that I don't yawn out loud.

On the side of the screen is a list, like the menu on the Emirate of Toys site, which I used by myself last year for my Christmas wants. The baby list is not very long. Babies only come in about six colours — we're getting one that matches Mother and me. Humans are a lot less interesting than Legos or iBots.

Nanny reads me all the diseases you can ask your baby not to have. Most of them are options, she says, which means we have to pay more. But I think we should pick them all, because a sick sister is not a good

thing. Angela Xhobi's sister has asthma, because she was made the old-fashioned way, without a menu, and she gets all the attention. I wouldn't like that at all.

Nanny takes a breath for another lecture, but I am saved when the iVid sings the Phone Call Song. Nanny sighs again and when she says, "Connect," I see that it's her mother, who calls every afternoon. Mrs Nanny is quite deaf, even with her implants, so Nanny taps SAVE on the baby screen and goes downstairs where she can shout without me hearing all the words. "Little pitchers," she says to her mother as she greys the upstairs iVid. I don't know what that means.

I slump back into my chair, because Nanny isn't here to tell me not to, and because she will be gone a long time. Her mother always has a lot to say. I stare at all the diseases, and then I see a better word at the bottom of the screen. PETS.

We don't harbour animals, because Nanny is allergic. (She was made the old-fashioned way, too.) But I'd like to see what we could have. I touch the screen to scroll down for more pets, and a Bubble Man appears, to tell me about a special offer. His picture seems to come out of the wall and stand right in front of me.

"Jellyfish DNA on sale," the Bubble Man says. He takes off his top hat, pulls a rabbit out of it, and holds it out towards me. The rabbit's fur glows a soft, bright green.

"Wow," I say.

"Bioluminescence, 50% off. Today only. Touch Box 306a to order!" He steps back into the screen and disappears with a little picture of smoke.

It only takes me a minute to find Box 306a and tap it to red. Then I SAVE and scroll back up to the disease boxes. It is good to leave things just the way you found them.

I sit very straight in my chair, humming, because I know a secret. Once I have my baby sister, I will never need my night-light again.

Nanny will be so proud.

Ellen Klages won a Nebula Award in 2005 for her story *Basement Magic*. Her first novel, *The Green Glass Sea*, will be published by Viking in October.

JACEY